Vol. 50 AUGUST 1959 No. 4

The Journal of Educational Psychology

CONTENTS

The Prediction of College Grades from the California Psychological Inventory and Scholastic Aptitude Test 135
JOHN L. HOLLAND

The Relationship of Intelligence and Achievement to Birth Order, Sex of Sibling, and Age Interval 143
SARAH M. SCHOONOVER

Creativity and the Self-Attitudes and Sociability of High School Students 147
LEANNE GREEN RIVLIN

Rational and Mathematical Relationships of Six Scoring Procedures Applicable to Three-Choice Items 153
JACK C. MERWIN

Interpretation of Differences among Averages and Individual Test Scores 162
FREDERICK B. DAVIS

Effects of "Brainstorming" Instructions on Creative Problem Solving by Trained and Untrained Subjects 171
SIDNEY J. PARNES AND ARNOLD MEADOW

Traditional Jewish Cultural Values and Performance on the Wechsler Tests 177
BORIS M. LEVINSON

AMERICAN PSYCHOLOGICAL ASSOCIATION

The Journal of Educational Psychology

RAYMOND G. KUHLEN, Editor
Syracuse University

Advisory Editors

LEE J. CRONBACH
University of Illinois
FRANCIS J. DIVESTA
Syracuse University
CHESTER W. HARRIS
University of Wisconsin
ERNEST R. HILGARD
Stanford University
NICHOLAS HOBBS
George Peabody College for Teachers
HAROLD E. JONES
University of California
IRVING LORGE
Teachers College, Columbia University
DOROTHEA MCCARTHY
Fordham University
T. ERNEST NEWLAND
University of Illinois
H. H. REMMERS
Purdue University
JULIAN C. STANLEY
University of Wisconsin
J. B. STROUD
State University of Iowa
DONALD E. SUPER
Teachers College, Columbia University
PERCIVAL M. SYMONDS
Teachers College, Columbia University
GEORGE G. THOMPSON
Syracuse University

The Journal of Educational Psychology publishes original investigations of problems of learning, teaching, and the psychological development of the individual. It contains articles on the following subjects: the psychology of school subjects; experimental studies of learning; the development of interests, attitudes, personality, social relations, and vocational orientation, particularly (but not exclusively) as related to school adjustment or the impact of the school program upon development and adjustment; emotion, motivation, and character; mental development; and methods. The last includes tests, statistical techniques, and research techniques in cross-sectional and developmental studies. Studies published deal with all levels of education—preschool, elementary, secondary, college, and adult education—and with all age groups.

Manuscripts and correspondence on editorial matters should be addressed to the Editor: Raymond G. Kuhlen, Department of Psychology, Syracuse University, 123 College Place, Syracuse 10, New York.

All manuscripts should be submitted *in duplicate* to facilitate referral to advisory editors and must be prepared to conform to the style and procedures described in the *Publication Manual of the American Psychological Association*. Because of the great demand for journal space, authors should give special attention to reducing the length of manuscripts while maintaining clarity of expression. Ordinarily, papers should not exceed 15 typed pages. Accepted manuscripts are published in order of their receipt. An author may, however, request early publication of accepted manuscripts if he is willing to assume the cost of printing the necessary extra pages.

ARTHUR C. HOFFMAN
Managing Editor

HELEN ORR
Promotion Manager

Communications including subscriptions, orders of back issues, and changes of address should be addressed to the American Psychological Association, 1333 Sixteenth Street, N.W., Washington 6, D. C. Address changes must reach the APA by the 10th of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Annual subscription: $8.00 (Foreign $8.50). Single copies, $1.50.

PUBLISHED BIMONTHLY BY:

THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.

Mt. Royal & Guilford Aves., Baltimore 2, Md.
and 1333 Sixteenth Street N.W., Washington 6, D. C.

Second-class postage paid at Baltimore, Md.
Copyright, 1959 by the American Psychological Association, Inc.

THE JOURNAL OF

EDUCATIONAL PSYCHOLOGY

Volume 50 August, 1959 Number 4

THE PREDICTION OF COLLEGE GRADES FROM THE CALIFORNIA PSYCHOLOGICAL INVENTORY AND THE SCHOLASTIC APTITUDE TEST[1]

JOHN L. HOLLAND

National Merit Scholarship Corporation, Evanston, Illinois

This study is designed to explore the usefulness of nonintellectual factors in predicting college grades and to provide needed information for the development of a theory of intellectual achievement. Since an earlier report indicates that the SAT and high school rank have only low efficiency for predicting the grades of a high aptitude sample within a narrow range of talent (Holland, 1958), the present study was performed to test the usefulness of the CPI and SAT, separately and in combination, as predictors of scholastic achievement for a sample of exceptionally talented college freshmen attending 291 colleges and universities. The results of this study provide a test of the predictive validities of these instruments, since the student sample was tested prior to college entrance.

Student Sample

The student sample was obtained from an earlier study (Holland, in press) in which National Merit finalists were tested with the CPI a month before the fall term of 1957. The SAT had been administered to this student sample about seven months before the fall term. The sample included 743 Merit Scholars and 578 Certificate of Merit winners drawn from a sample of 7500 finalists, the survivors of a nationwide competition in which 166,000 high school seniors participated (National Merit Scholarship Corporation, 1957). Because of their overall psychometric and demographic similarity, Merit Scholars and Certificate of Merit winners were combined, and the total sample was sorted randomly into two approximately equal subgroups. Boys and girls were analyzed separately in both subgroups. The first subgroup (standard sample) was used to develop regression equations for each sex, and the second subgroup (cross-validation sample) was used for cross-validation. The *N*s, means, and standard deviations for each sample on the CPI, SAT, and HPR variables are shown in Table 1.

Table 1 indicates that the standard and cross-validation samples for both sexes are comparable with respect to CPI, SAT, and HPR measures. Table 1 also shows the high aptitude levels of these students and the accompanying restriction in range. The male samples average 670 and 710 on the Verbal and Mathematical sections of the SAT. The female samples average 682 and 654 on these same variables. The corresponding standard deviations for these variables range from three fourths to less than

[1] This study was partially supported by research grants from the National Science Foundation and the Old Dominion Foundation. The author wishes to acknowledge the contributions of Donald L. Thistlethwaite and Laura Kent for their critical reviews of this paper.

TABLE 1

THE STANDARD AND CROSS-VALIDATION SAMPLES IN TERMS OF SCHOLASTIC APTITUDE TEST, CALIFORNIA PSYCHOLOGICAL INVENTORY, AND HONOR POINT RATIO

Scale	Boys				Girls			
	Standard (N = 476)		Cross-validation (N = 481)		Standard (N = 185)		Cross-validation (N = 179)	
	x	SD	x	SD	x	SD	x	SD
SAT-Verbal (SAT-V)	670.8	45.3	668.7	44.2	678.5	45.4	684.9	40.9
SAT-Math (SAT-M)	711.5	72.7	709.1	67.3	649.2	74.3	657.9	81.4
Dominance (Do)	31.0	5.9	31.1	5.8	30.5	5.4	30.6	5.8
Capacity for Status (Cs)	22.2	3.3	22.3	3.1	23.2	3.2	23.1	3.2
Sociability (Sy)	26.7	4.9	26.4	4.8	26.7	4.5	26.6	5.4
Social Presence (Sp)	37.8	6.2	37.8	5.7	36.7	5.3	36.1	5.7
Self-acceptance (Sa)	23.2	3.9	22.9	3.9	22.5	3.6	22.4	3.8
Sense of Well-being (Wb)	37.6	3.6	37.4	3.9	37.6	3.7	36.8	4.6
Responsibility (Re)	34.1	4.2	34.1	4.0	35.5	4.1	35.5	3.6
Socialization (So)	38.7	5.2	39.0	5.2	40.5	5.5	39.6	5.1
Self-control (Sc)	28.1	7.1	28.7	7.4	30.6	7.3	29.2	7.3
Tolerance (To)	25.6	3.8	25.8	3.7	26.9	3.0	26.2	3.7
Good Impression (Gi)	17.4	5.7	17.9	5.7	18.7	5.9	17.3	6.0
Communality (Cm)	25.7	1.8	25.7	1.8	25.8	1.7	25.7	1.9
Achievement via Conformance (Ac)	29.1	4.0	28.9	4.0	29.4	3.8	29.1	4.4
Achievement via Independence (Ai)	23.3	3.2	23.2	3.2	23.7	3.0	23.2	3.0
Intellectual Efficiency (Ie)	43.8	3.5	43.9	3.6	44.6	3.2	43.6	4.4
Psychological-mindedness (Py)	13.5	2.4	13.4	2.4	13.1	2.5	12.9	2.5
Flexibility (Fx)	11.9	3.9	11.8	3.9	11.7	3.8	11.8	3.7
Femininity (Fe)	16.7	3.9	16.8	3.8	22.8	3.5	22.9	3.1
Honor Point Ratio (HPR)	3.11	.61	2.97	.70	3.18	.55	3.14	.62

one half of the standard deviation obtained in the standardization of the SAT.[2]

CRITERION

Freshman grades in college, or honor point ratio, were used as the criterion of scholastic achievement. The grading systems of all the colleges in the study were converted to HPR by means of the standard formula[3] used by the majority of institutions. Generally, this formula was applied to collegiate grading systems by using the equivalences given in college regulations, or in letters from the colleges. In a few instances, numerical values had to be assigned to letter grades on the basis of the investigator's best judgment.

RESULTS

College grades were intercorrelated with the CPI and SAT variables for student samples which were classified in three different ways. First, zero-order correlations were computed for both the standard and cross-validation samples, and multiple regression equations derived from the standard samples were applied to the cross-validation samples. Second, both the standard and cross-validation samples were dichotomized as science or nonscience majors, and correlations were computed for these

[2] The SAT is standardized with a mean of 500 and a standard deviation of 100.

[3] All grades were converted by an honor point ratio formula where A = 4, B = 3, C = 2, D = 1, F = 0; grades were multiplied by credits per course and divided by total credits carried.

TABLE 2

THE CORRELATION OF GRADES (HPR) WITH THE SCHOLASTIC APTITUDE TEST AND THE CALIFORNIA PSYCHOLOGICAL INVENTORY ACROSS 291 COLLEGES

Scale	Boys		Girls	
	Standard	Cross-validation	Standard	Cross-validation
	(N = 476)	(N = 481)	(N = 185)	(N = 179)
SAT-Verbal	.12**	.04	.26**	.06
SAT-Math	.17**	.15**	.15*	.00
Dominance	−.06	−.02	−.01	.04
Capacity for Status	−.19**	−.11*	−.12	−.15*
Sociability	−.21**	−.12**	−.04	−.10
Social Presence	−.26**	−.18**	−.19**	−.24**
Self-acceptance	−.15**	−.12**	−.15*	.00
Sense of Well-being	−.07	.09*	.09	.01
Responsibility	.16**	.19**	.25**	.11
Socialization	.22**	.29**	.22**	.27**
Self-Control	.15**	.24**	.22**	.12
Tolerance	−.05	.05	.10	.02
Good Impression	.02	.19**	.03	.04
Communality	.02	−.02	.08	.07
Achievement via Conformance	.07	.19**	.28**	.09
Achievement via Independence	.05	.05	.18*	.05
Intellectual Efficiency	−.04	−.06	.12	−.05
Psychological-mindedness	.01	.08	.09	−.14
Flexibility	−.10*	−.14**	−.08	−.22**
Femininity	.26**	.21**	.19**	.14

* Significant at .05 level.
** Significant at .01 level.

subgroups. Third, correlations between grades and test variables were computed for eight colleges in which 24 or more of the students in the sample were enrolled. The results of these analyses are shown in Tables 2, 3, and 4.

Table 2 presents the correlations for grades against the CPI and SAT variables. To test the efficiency of prediction, multiple regression equations were derived from the standard male and female samples and applied to the corresponding cross-validation samples. To reduce computational labor, a sample of test variables was selected for inclusion in the regression equations. For the standard samples, we selected the single variable having the highest significant correlation from the SAT and from each of the four classes of CPI scales suggested by Gough (1957, pp. 12–13). Multiple correlations and regression equations were obtained by following the DuBois (1957) procedure.

In z-score form, the resulting equation for boys is: $z_o = .16M + .11So - .19Sp + .17Fe$. The multiple correlation for the standard male sample is .38, which is significant beyond the .01 level. Application of the equation to the male cross-validation group resulted in a correlation of .32 ($P < .01$), or only slight shrinkage.

The equation for girls is: $z_o = .25V - .14Sp + .06Re + .20Ac + .08Fe$. The multiple correlation for the standard female sample is .42, which is significant beyond the .01 level. The application of the equation to the female cross-validation sample resulted in a correlation of .23,

which is significant beyond the .01 level, although the cross-validation R is about one-half the standard R.

The zero-order correlations in Table 2 are also of interest. The mathematics factor of the SAT is statistically significant for both male samples (r = .17 and .15), but for only one of the female samples. The verbal factor is significant only for the male and female standard samples.

For boys, the Capacity for Status (Cs), Sociability (Sy), Social Presence (Sp), Self-acceptance (Sa), Responsibility (Re), Socialization (So), Self-control (Sc), Flexibility (Fx), and Femininity (Fe) scales of the CPI are statistically significant for both the standard and cross-validation samples. The Socialization scale has the highest pair of correlations, .22 and .29, for these samples. For girls, only the Social Presence (Sp), and Socialization (So) scales are significant in both samples. The small size of the female samples may account for the limited number of significant correlations on both instruments.

Table 3 summarizes the results of classifying students in the four standard and cross-validation samples as science or nonscience majors and recomputing the correlations between grades and test variables. This classification was used to determine what effect a more homogeneous grouping

TABLE 3

THE CORRELATION OF GRADES WITH THE SCHOLASTIC APTITUDE TEST AND CALIFORNIA PSYCHOLOGICAL INVENTORY FOR SCIENCE AND NONSCIENCE SAMPLES

Scale	Boys				Girls			
	Standard		Cross-validation		Standard		Cross-validation	
	Science (N = 345)	Non-science (N = 95)	Science (N = 354)	Non-science (N = 84)	Science (N = 83)	Non-science (N = 79)	Science (N = 95)	Non-science (N = 70)
SAT-Verbal	.12*	.12	.02	.04	.17	.32**	.12	−.04
SAT-Math	.22**	.09	.16**	.06	.10	.06	.06	.08
Dominance	−.11*	−.03	−.02	−.10	.09	.00	.11	−.01
Capacity for Status	−.19**	−.27**	−.06	−.28**	−.13	−.03	−.12	−.14
Sociability	−.22**	−.18	−.11*	−.23*	−.06	−.06	−.05	−.05
Social Presence	−.23**	−.37**	−.16**	−.25*	−.18	−.07	−.19	−.32**
Self-acceptance	−.17**	−.14	−.10	−.33**	−.06	−.16	.07	−.11
Sense of Well-being	.02	−.30**	.11*	.09	.16	.13	.08	.02
Responsibility	.13*	.21*	.17**	.23*	.25*	.16	.21	−.02
Socialization	.24**	.18	.31**	.29**	.26*	.23*	.28**	.30**
Self-control	.18**	.05	.24**	.25*	.20	.19	.21	.04
Tolerance	−.01	−.20	.04	.05	.02	.12	.16	−.08
Good Impression	.03	−.03	.23**	−.01	.10	−.06	.15	−.07
Communality	.02	−.06	−.02	−.08	.06	.10	.13	−.01
Achievement via Conformance	.08	.02	.19**	.23*	.32**	.16	.25*	.01
Achievement via Independence	.08	−.16	.03	.11	.15	.26*	.15	−.03
Intellectual Efficiency	.01	−.15	−.09	.09	.14	.04	.10	−.14
Psychological-mindedness	.06	−.14	.13*	−.03	.17	.01	−.05	−.19
Flexibility	−.09	−.13	−.13*	−.16	−.09	−.03	−.24*	−.25*
Femininity	.25**	.32**	.21**	.09	.16	.20	.10	.18

Note.—The Ns for the standard and prediction samples have been reduced by eliminating students whose choice of major field is unkown or ambiguous.

* Significant beyond .05 level.

** Significant beyond .01 level.

TABLE 4

THE CORRELATION OF GRADES (HPR) WITH SCHOLASTIC APTITUDE TEST AND CALIFORNIA PSYCHOLOGICAL INVENTORY VARIABLES WITHIN COLLEGES

Scale	Boys						Girls	
	C.I.T. (N = 26)	Harvard (N = 81)	M.I.T. (N = 73)	Princeton (N = 44)	Stanford (N = 27)	Yale (N = 49)	Radcliffe (N = 24)	Wellesley (N = 24)
SAT-V	.09	.15	.22	.36*	.18	.29*	.49*	.40*
SAT-M	.16	.07	.49**	.31*	.23	.14	.31	.29
Dominance	.33	−.16	.02	.10	−.33	−.04	−.30	−.01
Capacity for Status	.19	−.30**	.00	−.03	−.28	.16	−.19	−.05
Sociability	.01	−.16	.00	−.06	−.35	−.12	−.03	−.04
Social Presence	−.22	−.33**	.01	−.03	−.47*	−.22	−.33	−.05
Self-acceptance	−.07	−.32**	−.03	.17	−.16	−.13	−.25	.29
Sense of Well-being	.22	.17	.23*	.02	.12	.17	.01	.34
Responsibility	.41*	.30**	.10	.37*	.24	.32*	.33	.12
Socialization	.57**	.29**	.23*	.35*	.33	.23	.31	.25
Self-control	.45*	.31**	.29*	.23	.44*	.32*	.22	.28
Tolerance	.21	−.05	.19	−.10	.13	−.07	.16	−.05
Good Impression	.41*	.08	.33**	.22	.30	.31*	.03	.01
Communality	.10	−.01	−.03	−.09	−.14	.03	.04	.15
Achievement via Conformance	.51**	.23*	.35**	.21	.32	.18	.23	.24
Achievement via Independence	−.14	.08	.17	.25	.43*	.13	.37	−.16
Intellectual Efficiency	.18	−.05	−.03	.24	.21	−.06	.32	−.04
Psychological-mindedness	.05	.15	.21	−.49**	.04	.20	.43*	−.15
Flexibility	−.32	−.05	−.09	.00	−.10	−.07	.01	−.19
Femininity	.37	.33**	.20	.54**	.51**	.30*	.41*	−.39

Note.—C.I.T. and M.I.T. stand for California Institute of Technology and Massachusetts Institute of Technology, respectively.

* Significant beyond .05 level.

** Significant beyond .01 level.

would have on the correlations between grades and test variables and to discover if academic achievement in science requires a different set of aptitudes and personality variables than achievement in nonscience.

Since the average of the absolute values of the 80 correlations in Table 3 is .13, which closely approximates the average r of .12 for Table 2, it is clear that this classification produced no change in the level of relationships between grades and test variables. Table 3 also suggests that, for the freshman year, 18 of the 20 variables predict grades for science and nonscience majors with about equal efficiency. Only the Mathematics factor for boys and the Ac scale for girls are unique, in that these variables are significantly correlated with grades for both standard and cross-validation samples, but neither is significantly related to grades for the four nonscience samples.

The correlation of SAT and CPI variables within individual colleges is shown in Table 4. This analysis was performed to evaluate the effect of a more reliable criterion (the grading system of a single college rather than the equating of multiple systems) and to estimate the range of predictive validities for single test variables from college to college. The 80 correlations for the eight colleges included in Table 4 average .21. In contrast, the 80 correlations across colleges shown in Table 2 average .12. Presumably this difference is due to differences in grading systems as well as in the aptitude level of student populations.

The general pattern of correlations in Ta-

ble 4 resembles that found for the total sample; that is, high grades are negatively associated with the Cs, Sy, Sp, Sa, and Fx scales and positively associated with the Wb, Re, So, Sc, Gi, Ac, Ai, Py, and Fe scales. In scale terms, these findings suggest that the high achiever lacks capacity for status, is unsociable, lacks poise and self-confidence, is self-deprecating and inflexible, minimizes worries and complaints, is conscientious and responsible, is well controlled, and creates a favorable impression, does well academically under direction but is not as adept in situations demanding independent judgment, is interested in and responsive to the feelings of others, and has feminine interests. In contrast, the low achiever is poised and socially skillful, has positive self-attitudes, is flexible, admits worries and complaints, has less intense superego qualities, is impulsive, creates a less favorable impression, possesses less motivation for academic achievement, and has more extraceptive and masculine interests. Although individual colleges follow this general pattern, the eight colleges show a wide range of differences on a more limited number of scales. These findings suggest that achievement in the majority of colleges results from a general cluster of personality and aptitude variables, but that a given college may demand, in addition, a limited number of specific characteristics.

For example, note that the correlations between SAT-V and grades range from .09 to .49; similarly, the correlations for SAT-M range from .07 to .49. In addition, the relative predictive value of V and M vary markedly within a given college. The Dominance scale ranges from .30 to −.33; the Capacity for Status scale ranges from .19 to −.30; the Social Presence scale ranges from .01 to −.47; the Self-acceptance scale ranges from .29 to −.32; the Good Impression scale ranges from .01 to .41; the Achievement via Independence scale ranges from −.14 to .43; the Psychological-mindedness scale ranges from .43 to −.49; and the Femininity scale ranges from .54 to −.39.

The patterning of predictors within and between colleges may also be due to the variation in institutional environments. Thistlethwaite (in press) found college press to be related to student achievement. Consequently, the interaction of student motivational and personality characteristics and college presses may in part produce the correlational differences among colleges shown in Table 4. This interpretation appears plausible since the majority of standard deviations for the individual colleges listed in Table 4 are smaller than standard deviations for the total samples shown in Table 2. Accordingly, the greater average size of the Table 4 correlations compared to the Table 2 correlations is probably due to the interaction of students and college environments as well as the use of a more reliable criterion rather than to differences in student homogeneity on the test variables.

Taken together, the results suggest that for some individual colleges, specific formulae which would be substantially effective in predicting grades could be developed. This assumption appears especially plausible since the multiple *R*s attained here were accomplished across a large number of colleges with a sample having a narrow range of talent. Further, the zero-order correlations for individual colleges shown in Table 4 are substantially higher than those obtained across colleges in Table 2.

Discussion

For the SAT, the Verbal and Math factors appear to be about equally efficient in predicting grades. For the CPI, the Sp, So, Re, Ac, and Fe scales have useful predictive validity, both alone and in combination with the SAT. The Sp and So scales appear more efficient, since they are significantly related to grades for all four total samples across colleges. The significant correlation of the CPI So scale with grades in all four major samples supports Gough's (1955) theory that achievement

and underachievement among gifted persons is a specific facet of the general problem of socialization.

These results generally support Gough's (1955, 1957) findings, but they fail to confirm the validity of the Ai and Ie scales. However, the high aptitude levels and outstanding high school achievement records of this sample may explain these unexpected variations, particularly the relative failure of the Ai and Ie scales and the unexpected predictive efficiency of the Social Presence, Capacity for Status, and Flexibility scales. When the mean scale scores for the four across-college samples are profiled, using high school norms, the two highest scores for both male and female samples occur on the Ai and Ie scales, which are elevated approximately 2 standard deviations above the mean. Since the Scholars and Certificate winners have unusual high school records (the average HSR is about 99 for Scholars and 95 for Near-winners), it is significant that these students should peak on these scales. The homogeneity of scores on these scales may also reduce their effectiveness, so that other CPI scales are more predictive in this study (see Table 1).

When the CPI and SAT are compared, the CPI seems generally more efficient; however, the comparison is an ambiguous one because of the differences in time of testing and the circumstances of the administration of these instruments. Moreover, the Scholarship Qualifying Test was used as a screening device to select a final pool of 7500 students from the 166,000 participants in the National Merit program. As a basis for awarding scholarships, the SAT was used to retest this final pool of students. Since the SAT and SQT are similar in content and construction, the selection of the final pool by the SQT probably attenuates the correlations obtained for the SAT. The CPI was not used as a screening device, a fact which was known to the students in the sample. It seems likely that selection on the basis of SQT scores attenuated correlations for the CPI less than correlations for the SAT.

Probably more important than the question of the relative efficiency of these tests is the finding that at a high level of scholastic aptitude, personality variables may yield validity coefficients which are two to almost three times as great as those obtained using aptitude measures alone. Whether this would be true if both instruments were used as selective devices is, of course, not known. The present evidence contrasts sharply to that of most academic prediction studies, which report that the addition of a personality variable usually adds only a relatively small increment to the aptitude-grades correlation if it increases the correlation at all. Using a combination of high school rank and SAT to predict grades is also relatively inefficient for samples of the present aptitude level. In a similar study of National Merit Scholars in 1956 (Holland, 1958), it was found that SAT-V, SAT-M, and HSR produced a multiple R, non-cross-validated, of only .16 for a total male sample of 394 science majors only; this result is only one-half the cross-validated R of .32 for the 1957 male sample.

Summary

The CPI and SAT are useful in predicting college freshman grades for a sample of high aptitude high school seniors. Multiple regression equations for the CPI and SAT in combination cross-validate, and the resulting multiple Rs are two to three times as great as the zero-order r's obtained for the SAT alone. Although CPI variables are generally more effective than the SAT variables in the present study, the individual scales in these instruments show wide variation in validity from college to college.

REFERENCES

DuBois, P. H. *Multivariate correlational analysis*. New York: Harper, 1957.

Gough, H. G. Factors related to differential achievement among gifted persons. Paper read at APA, San Francisco, 1955.

GOUGH, H. G. *Manual for the California Psychological Inventory*. Palo Alto: Consulting Psychologists Press, 1957.

HOLLAND, J. L. Prediction of scholastic success for a high aptitude sample. *Sch. Soc.*, 1958, **86**, 290–293.

HOLLAND, J. L. Determinants of college choice. *Coll. & Univer.*, in press.

NATIONAL MERIT SCHOLARSHIP CORPORATION. *Second annual report*. Evanston: Author, 1957.

THISTLETHWAITE, D. L. The development of talent by American colleges. *Science*, in press.

Received January 26, 1959.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 4, 1959

THE RELATIONSHIP OF INTELLIGENCE AND ACHIEVEMENT TO BIRTH ORDER, SEX OF SIBLING, AND AGE INTERVAL[1]

SARAH M. SCHOONOVER

Northwest Guidance Center, Lima, Ohio

This investigation is a supplement to the writer's previous study of resemblances in the mental and educational ages of siblings (Schoonover, 1956). In the original study, three different methods of analysis of longitudinal growth records, i.e., means of the average differences, percentage reduction of difference by family membership, and correlation produced evidence of a substantial amount of sibling resemblance in intelligence and achievement. Resemblances in intelligence were somewhat greater than they were in achievement.

The purpose of this study is to investigate the following questions: (*a*) What is the relationship of the ordinal position of a child to his mental test performance? (*b*) What is the relationship of the sex of a child's sibling to his mental test performance? (*c*) What is the relationship between age interval and degree of resemblance in mental achievement of siblings?

Studies involving familial resemblances in mental ability have been numerous during the past 90 years. The contribution of the above investigations is that they employ the longitudinal approach to this problem in a manner that has not been utilized with data analyzing sibling resemblances in intelligence or achievement.

SIBLINGS SELECTED FOR STUDY

The data were obtained from the records of the University Elementary School at the University of Michigan. All true sibling pairs, with chronological age overlap, and with four or more scores per sib on the Stanford-Binet test and on the Stanford Achievement test, from the fall of 1929 through the spring of 1951, were used in this study. With these qualifications, 59 sibling pairs were found for intelligence; 64 pairs for arithmetic, education, reading, and spelling; 42 pairs for literature and social studies; 40 pairs for language; and 38 pairs for science. For additional details regarding description and selection of data the reader is referred to the earlier report (Schoonover, 1956).

METHODS FOR THE COMPARISON OF LONGITUDINAL SIBLING RECORDS

For each family included in this study a mental growth graph and eight achievement growth graphs were constructed. This meant that 344 growth graphs and 757 individual growth curves were plotted.

The linear equation best fitting the data was found to eliminate the observed variation and to determine some constant rate of growth which may be used to characterize the observed results. The equation of a straight line, $y = ax + b$, which gives the slope and the intercept of the line used to describe the growth-age relationship, was found by the method of the least squares fit. The linear fit for each child for intelligence and for each of the eight achievement variables was plotted graphically.

For each pair of siblings the limits of the overlap of their chronological ages were found, and from these the midpoint of the age overlap was computed. From the linear best fit, the age scores of each sibling were read at these midpoints. By subtracting the score equal to the midpoint of the age overlap from each sibling's chronological age score, the average difference for the

[1] The writer is indebted to Willard C. Olson and to Byron O. Hughes for their criticisms and suggestions in the preparation of this manuscript.

overlap period was found. In other words, each child's mental and/or achievement age score difference from the theoretical norm was obtained. The means of these differences were calculated for the older siblings and for the younger siblings.

Using the method described above, the means of the differences between mental and/or achievement age scores and midpoints of the chronological age overlap were calculated for sibs with brothers and for sibs with sisters. To ascertain the reliability of the differences between the sib-pair means, t tests were employed.

The chronological age difference in months and the midpoint of the age overlap was found for each sibling pair. From the linear best fit, the age scores of each sibling were read at these midpoints. The average difference for each sibling pair was found by subtracting the smaller age score from the larger age score. Correlation coefficients were computed to discover the relationship between the chronological age difference and the average score difference for the sibling pairs. A Pearson product-moment formula was used, as was a t test of an observed correlation.

Findings

Older and younger siblings. The results secured by the comparison of the means of the average differences from chronological age norms for older sibs and for younger sibs are given in Table 1 for intelligence and achievement.

TABLE 1

Means of Average Differences From Chronological Age Norms for Older and Younger Siblings, Expressed in Months

Mental or Achievement Age Measure	Older Sib	Younger Sib
Mental	24.8	25.7
Arithmetic	1.0	1.4
Education	9.4	8.0
Language	35.8	35.7
Literature	23.9	21.7
Reading	18.1	14.2
Science	25.1	24.2
Social studies	13.4	13.3
Spelling	−1.1	−1.4

These older and younger siblings consistently were found to have means of average differences that were very similar to each other. No differences were found to be reliable at the .05 level.

These findings are in accord with those of Hsiao (1931), Jones and Hsiao (1928), and Griffitts (1926). They are incongruent with those of Thurstone and Jenkins (1929), Willis (1924), and Koch (1954). Jones (1954) has pointed out that in normal samples, when methodological difficulties are adequately controlled, no birth order differences in intelligence occur. The findings of this study not only support this viewpoint but, in addition, reveal no birth order differences in achievement.

Siblings with brothers and siblings with sisters. The results obtained by the comparison of the means of the average differences from chronological age norms for sibs with brothers and for sibs with sisters are shown in Table 2.

TABLE 2

Means of Average Differences From Chronological Age Norms for Siblings With Brothers and Siblings With Sisters, Expressed in Months

Mental or Achievement Age Measure	Sibs with Brothers	Sibs with Sisters
Mental	26.2	23.9
Arithmetic	3.0	−0.9
Education	10.3	6.5
Language	44.3	25.8
Literature	27.6	17.1
Reading	17.9	13.4
Science	32.7	17.1
Social studies	19.5	6.7
Spelling	1.2	−4.1

Siblings with brothers consistently were found to have larger means of average differences than siblings with sisters. In other words, sibs with brothers had higher scores than sibs with sisters in all mental and achievement measures. The differences between these sib means were significant

at the .01 level or lower in language, literature, science, and social studies, and at the .05 level in arithmetic. The sib mean differences in intelligence, reading, education, and spelling were not significant at the .05 level.

These results substantiate those found by Koch (1954) and are in disagreement with those secured by Tabah and Sutter (1954). As an explanation of the finding that the possession of a male sib appears to influence test performance, Koch has pointed out that possibly the more aggressive, vigorous, and competitive male alerts his sib to a greater extent than does the more passive female.

Age interval and average score difference for siblings. The results secured by the method of correlation coefficients for chronological age interval and average score difference for sibling pairs are given in Table 3. No relationship was found between age interval and average score difference for siblings, with correlations ranging from −0.09 to +0.18 for the nine mental and achievement measures.

It may be observed that these correlations substantiate each other and give some indication of the value of the correlations in the population, since they are all in a relatively narrow range.

These findings are harmonious with those of Conrad (1931), Finch (1933), and Richardson (1936). On the basis of the above results, it appears that variations in birth intervals are without influence on measures of intelligence and achievement.

As a suggestion for further research it may be pointed out that if a large number of comparable longitudinal sibling records could be secured, the interactions of the variables, i.e., ordinal position, sibling's sex, and spacing could be investigated.

TABLE 3

CORRELATIONS BETWEEN BIRTH INTERVAL AND DIFFERENCES BETWEEN SIBLINGS IN MENTAL AND ACHIEVEMENT AGES

Mental or Achievement Age Measure	Corelation Coefficient
Mental	0.03
Arithmetic	0.18
Education	0.00
Language	0.11
Literature	0.01
Reading	−0.09
Science	−0.07
Social studies	0.03
Spelling	−0.07

SUMMARY

An analysis was made of sibling performance on intelligence and achievement tests utilizing longitudinal data. (*a*) No significant differences were found between older and younger siblings in intelligence or achievement as measured by deviation from the norms for chronological age. Thus priority of birth in a family gave no advantage in intelligence or achievement. (*b*) Sibs, irrespective of sex, with brothers consistently had higher mental and achievement ages than sibs with sisters. (*c*) The relationship between interval between births and the average difference in intelligence and achievement for sibling pairs was insignificant.

REFERENCES

CONRAD, H. S. Sibling resemblance and the inheritance of intelligence. Unpublished doctoral dissertation, Univer. of California, 1931.

FINCH, F. H. A study of the relation of age interval to degree of resemblance of siblings in intelligence. *J. genet. Psychol.*, 1933, **43**, 389–404.

GRIFFITTS, C. H. The influence of family on school marks. *Sch. Soc.*, 1926, **24**, 713–716.

HSIAO, H. H. The status of the first born with special reference to intelligence. *Genet. psychol. Monogr.*, 1931, **9**, 1–118.

JONES, H. E., & HSIAO, H. H. A preliminary study of intelligence as a function of birth order. *J. genet. Psychol.*, 1928, **35**, 428–433.

JONES, H. E. The environment and mental development. In L. Carmichael (Ed.), *Manual of child psychology*. New York: Wiley, 1954.

KOCH, HELEN L. The relation of 'primary mental abilities' in five- and six-year-

olds to sex of child and characteristics of his sibling. *Child Develpm.*, 1954, **25**, 209–223.

RICHARDSON, S. K. The correlation of intelligence quotients of siblings of the same chronological age levels. *J. juv. Res.*, 1936, **20**, 186–198.

SCHOONOVER, SARAH M. A longitudinal study of sibling resemblances in intelligence and achievement. *J. educ. Psychol.*, 1956, **47**, 436–442.

TABAH, L., & SUTTER, J. Le niveau intellectuel des enfants d'une meme famille. *Ann. Eugen. Lond.*, 1954, **19**, 120–150.

THURSTONE, L. L., & JENKINS, R. L. Birth order and intelligence. *J. educ. Psychol.*, 1929, **20**, 641–651.

WILLIS, C. B. The effects of primogeniture on intellectual capacity. *J. abnorm. soc. Psychol.*, 1924, **18**, 375–377.

Received July 1, 1958.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 4, 1959

CREATIVITY AND THE SELF-ATTITUDES AND SOCIABILITY OF HIGH SCHOOL STUDENTS[1]

LEANNE GREEN RIVLIN

Brooklyn College

Current concern regarding the adequacy of our scientific and technical manpower has stimulated considerable interest in the area of the intellectually gifted. The pressing need to identify and encourage those students who are capable of high-level performance has directed attention to the techniques for recognizing the capable individual as early in his school career as possible.

A study of creativity of the intellectually gifted is one approach to the larger problem. Creativity as a focus emphasizes the originality aspect of giftedness. The limited research on creativity has approached the study from three views: the creative process, the creative product, and the creative personality. The creative act appears to be the result of a complex interaction, in which a number of necessary conditions have been met. These conditions include:

1. A basic minimum of intelligence (one that is considerably higher than average)
2. Skill in a particular area (called "talent," "technique," or "technical ability")
3. Training and/or experience in the area in which there is skill
4. An opportunity to utilize or actualize 1, 2, and 3
5. The personality characteristics that facilitate the functioning of 1 through 4

The factor-analytic studies of Guilford and his associates (Wilson, Guilford, & Christensen, 1953) might be regarded as an investigation of the process and product aspect of creativity. Gough (1955), Stein and Meer (Stein, 1953; Stein & Meer, 1954), Barron (1953), and Drevdahl (1956) have all examined personality factors relevant to creative effort. These studies indicate an independence or self-assertive factor as characteristic of creative individuals. However, the act of originality is popularly associated with extreme social isolation, intense and private self-searching: the picture of the artist shut up in his garret, the withdrawn writer, the scientist wrapped up in his world of formulae and test tubes. Little objective evidence for the reality of the traits of isolated social maladjustment has been offered other than occasional life histories and biographies of some creative individuals.

The major question investigated in this study is whether high school students selected as creative and equally able high school students selected as noncreative differ with regard to (*a*) self-attitudes, (*b*) sociability. The relation to creativity of intelligence, socioeconomic status, parental occupation, and parental education was also studied.

METHOD

The initial task involved the development of a technique for the identification of the creative and noncreative students. In view of the variability of the expression of creativity, it seemed advisable to employ a procedure based upon extended observation of the students involved. Nomination by teachers was adopted as the procedure most likely to take this variability into account. The criteria of creativity used by the teachers when making the nominations were developed in a preliminary study. A

[1] This paper is adapted from a dissertation submitted in partial fulfillment of the requirements for the degree of doctor of philosophy under the Joint Committee on Graduate Instruction, Columbia University. The author wishes to thank Arthur T. Jersild, Millie Almy, Irving Lorge, and Wayne Dennis for their encouragement and assistance.

critical incident (Flanagan, 1954) form was completed by 225 teachers from high schools throughout New York City, and the analysis revealed 14 criteria of creativity. In summary, these were:

1. Gives work a "personal touch"
2. Will venture into unfamiliar or new areas
3. Is sensitive to the potentialities of media
4. Works with enthusiasm and pleasure
5. Demonstrates judgment on the basis of personal standards, a sense of appropriateness, and taste
6. Is able to "let himself go" and freely respond to the source of stimulation
7. Displays a capacity for self-direction and independence
8. Simplifies a complex task by perceiving and emphasizing the essentials
9. Uses past knowledge to interpret a present problem in a manner that is original and meaningful
10. Displays judgment and foresight in planning work
11. Questions and tests the implications of facts
12. Understands when to give up plans or ideas that seem impractical or inadequate
13. Flexible in approach to problems
14. Demonstrates imaginative and original solutions to problems

The nominating teacher attended a training conference during which the criteria were discussed, and each was then directed to select five intelligent, creative students and five equally able noncreative students in his classes on the basis of the criteria. If he could not nominate five that would appropriately fit into each category, he was permitted to select fewer.

The students selected were given a battery of tests. These included a modification of Bills' Index of Adjustment and Values (Bills, Vance, & McLean, 1951), and the Social Adjustment subtest of the Bell Adjustment Inventory. Two of Bills' three ratings were used in the present study. These were "concept of self" and "concept of ideal-self." The "acceptance of self" category was omitted, since previous testing revealed some difficulties on the part of the students in using this category.

The test form contained questions regarding parental education and occupation. Four sociometric questions were also included, in which the nominees were asked to select the students in their grade whom they would: (*a*) like to invite to a party, (*b*) like to belong to their school club, (*c*) like to work with on a class committee, (*d*) consider most creative. Pintner IQs and scores on the Iowa Tests of Educational Development were obtained from school records.

TABLE 1

DISTRIBUTION OF 10TH- AND 11TH-GRADE STUDENTS NOMINATED AS CREATIVE AND NONCREATIVE

Group	10th Grade	11th Grade
Creative boys	15	14
Creative girls	17	16
Noncreative boys	24	14
Noncreative girls	10	16
(Total N = 126)		

Subjects

The *Ss* were 126 10th- and 11th-grade New York City high school students nominated by 25 honor class teachers, art teachers, or music teachers as either creative or noncreative. These special classes were selected in an effort to obtain *Ss* who were relatively homogeneous in each interest area. The classes represented all academic subject areas.

Treatment of Results

The data were subjected to an analysis of variance. The 126 *Ss* were separated into subgroups according to creativity (creative or noncreative), sex, and grade (10th or 11th grade). This resulted in eight unequal cell subgroups, summarized in Table 1, which required an analysis of variance of cell means. Since responses to the sociometric questions were not available for the

entire 10th and 11th grades, one 10th-grade class was studied for a rough index of student evaluation of peer social attractiveness. This class, designated "4-H" had the highest number of nominees, with 17 students represented.

In all comparisons, the .05 level of confidence was adopted.

Results

The basic results, except for the sociometric data, are indicated in Table 2, containing means and standard deviations for the creative and noncreative students, with the sexes separate but the grades combined. Table 3 contains the results of the analysis of variance based on the cells described in Table 1.

Self-Attitudes

There were no significant differences in self-attitudes when the evaluations of creative and noncreative students were compared. On the "ideal-self" section the boys'

TABLE 2

Comparison of the Creative and Noncreative Boys and Girls (Grades Combined) as Evidenced in the Means and Standard Deviations of the Several Measures[a]

	Boys				Girls			
	Creative		Noncreative		Creative		Noncreative	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Self Concept	247.86	16.48	247.47	19.57	249.61	15.89	244.92	15.28
Ideal-Self	279.86	11.00	278.08	10.79	284.79	6.96	280.69	11.95
Self-Ideal discrepancy	38.90	16.59	37.42	18.40	41.27	17.61	44.46	14.34
Sociability (Bell subtest)	7.06	4.87	9.21	5.32	5.36	3.11	8.85	4.45
Iowa	24.17	4.28	20.21	4.06	20.54	5.06	19.50	3.28
IQ (Pintner)	140.21	15.45	127.97	14.19	128.03	16.24	127.54	13.82

[a] *N*s are indicated in Table 1.

TABLE 3

Analysis of Variance of the Sociability Scores, Iowa Scores, and IQs of Students Nominated as Creative and Noncreative

Sources of Variation	Sociability			Iowa		IQ	
	df	Mean Squares	*F*	Mean Squares	*F*	Mean Squares	*F*
Sex	1	60.47	3.26	158.60	7.70**	1134.02	4.75*
Grade	1	1.17	—	26.12	1.27	61.90	—
Creativity (Cr.)	1	277.18	14.94**	168.75	8.20**	1139.06	4.78*
Sex × Cr.	1	17.78	—	53.90	2.62	1098.64	4.61*
Cr. × Grade	1	18.24	—	19.98	—	38.44	—
Sex × Grade	1	18.97	1.02	.00	—	.36	—
Sex × Grade × Cr.	1	.56	—	16.66	—	5.63	—
Within cells	118	18.55		20.59		238.52	
Total	125						

* Significant at the .05 level.
** Significant at the .01 level.

TABLE 4

COMPARISON OF THE CREATIVE AND NONCREATIVE STUDENTS (GRADES COMBINED) AS EVIDENCED IN THE MEANS OF THE SOCIOMETRIC MEASURES[a]

Sociometric Question	Creative		Noncreative	
	M	*SD*	*M*	*SD*
(1) invite to party	1.60	1.58	.43	.53
(2) belong to club	1.60	1.43	.28	.50
(3) class committee	2.20	1.81	.28	.50
(4) creative	2.00	1.88	.14	.38

[a] Total *N* was 17, 10 creative *Ss* and 7 noncreative *Ss*, all tenth graders with sexes combined.

scores were significantly lower than those of the girls, representing a lower ideal.

An analysis of 17 of the adjectives was then undertaken to determine the possibility of internal differences. The adjectives were selected because they embodied qualities that Gough (1955) and Barron (1953) had indicated as discriminating individuals possessing traits relevant to creativity. This analysis revealed four adjectives which discriminated at the .05 level: creative, imaginative, popular, and shy. The creative *Ss* had more positive self-evaluations on these traits.

Sociability

On the Bell subtest, the creative students had significantly lower sociability scores (representing higher social confidence) than the noncreative students. Table 3 indicates this difference, significant at the .01 level. There was no significant difference between the two grades studied, nor was there a significant sex difference.

The number of votes received by each student on each question yielded the sociometric score. An analysis of the sociometric choices of the 17 creative and noncreative nominees from Class "4-H" revealed that the creative students were more popular, and also were selected more often as possessing creativity. These differences, which are evident in Table 4, were significant at the .05 level. The votes for the creative students came from both their creative and noncreative classmates.

Intellective Measures

Both the Iowa tests and the Pintner IQs revealed significant differences between the sexes and between the creative and noncreative students. The boys' Iowa scores were superior to those of the girls, with the difference significant at the .01 level, as indicated in Table 3. A similar set of differences appeared on the IQs, significant at the .05 level. The creativity difference on both the Iowa scores and IQs came from the boys, for the creative girls did not score very much higher than the noncreative girls. We might anticipate an interaction factor to emerge in the analysis of variance. It appeared on the IQs, although not on the Iowa scores, possibly due to the very slight tendency for the creative girls' Iowa scores to be higher than those of the noncreative girls. When the intellective measures of the nominated students were compared with the scores of all students in their respective grades, the scores of nominees, both creative and noncreative, were consistently higher than their grade averages. In the case of the IQs, the mean score of the nominees in each grade was 130. This was 24 points higher than the 10th-grade mean, and 16 points higher than that of the 11th grade.

Background Factors

The parental occupation level of the creative students did not differ significantly from that of the noncreative students. However, the creative students' parents reached a significantly higher level of education. The difference was significant at the .05 level. Parental education was represented by four scale points: 1 when elementary school was the last attended; 2 for high school; 3 for college; and 4 for

postgraduate school. The average score for all creative students' parents was 2.6, and that of the noncreative group's parents was 2.07.

Discussion

When the self-attitudes and sociability of students nominated as either creative or noncreative were studied, some of the common stereotypes reputed to characterize the creative personality were not substantiated. Recognizing procedural limitations (e.g., possible teacher bias in selection in spite of training; the limitations of paper-and-pencil tests, etc.) some group differences appear. Students nominated as creative did not rate themselves lower on self-attitudes than noncreative students as might be anticipated, in line with the common conceptions of the creative individual. However, the self-attitude test contained 60 separate traits, and the total score may obscure internal differences. This is suggested by the item analysis of 17 of the adjectives, some of which were found to differentiate in the direction of more positive self-evaluations on the part of the creative students. Some traits may be more vital to creativity than others. Perhaps the traits themselves may not be as significant as the individual's reaction to his self-evaluations. Almost all students, creative and noncreative, revealed some discrepancy between the way they saw themselves and the way they would like to be. The distinction between the creative and noncreative person may rest not in the score itself, but in the student's response to the self-estimate. The discrepancy may act to constrict or inhibit the creative productivity of the noncreative student, while it either stimulates or at least does not impede the creative individual's push ahead. It now seems essential to consider self-evaluations in a more sensitive manner, examining specific traits rather than an over-all score.

The greater social confidence and popularity of the student nominated as creative also does not support the common conception of the antisocial creator. However, interest in and capacity for social interaction do not preclude moments of intense isolation and concentration so necessary for creative work. Rather, the creative student appears to be capable of both concentration and socialization. It must be acknowledged that the teacher nomination itself may be a form of sociability index. Thus, we may be studying a group of sociable creators. Whether another group of quiet, seclusive creators also exists we cannot say on the basis of this study, although it suggests a direction for future research. With regard to the pronounced sociability differences, a number of possible explanations exist:

1. Lack of social confidence may inhibit the realization of a creative potential.
2. Lack of social confidence may prevent a functioning creativity from being noticed and shared by others.
3. The social confidence of the creative students may be a superficial skill, and not identical with sympathy and sensitivity.
4. The manifestation of creative endeavors itself may increase the individual's social confidence; the lack of special talents may diminish the potential, in one important area, for gaining social confidence.

The significant intellective differences between creative and noncreative boys (with negligible differences in the case of girls) and the greater number of boys placed in the noncreative group (particularly in the 10th grade) suggests either the existence of a nominator bias in the selection of the groups, or perhaps a creativity difference between the sexes. Further work with a more rigorous nomination procedure seems necessary. However, when we examine the discrepancy between the intelligence of creative and noncreative boys, it would seem improbable that this difference is the major factor determining creativity.

What some of the other factors are can only be suggested by this research. The background data raise the possibility that a family of somewhat higher educational

level may play a role in stimulating creativity. Further consideration of the intellectual climate of the home is indicated.

Summary and Conclusions

This study has attempted to examine the validity of certain assumptions regarding the creative personality in a high school population. One hundred and twenty-six *Ss* were nominated by their teachers as either creative or noncreative, and the two groups were compared with regard to self-attitudes, sociability, and certain background information. The results indicated that the student selected as creative emerged as a rather sociable individual evaluating himself as more confident in his relationships with people, more popular and creative as viewed by his peers than his noncreative counterpart. The creative student did not differ in over-all self-attitudes from the noncreative. It was suggested that a number of trait combinations might be equally conducive to creativity. The study does indicate that in the case of high school students two factors associated with creativity appear to be social confidence and parents who have attained a somewhat higher educational level. A number of directions for future research were suggested.

REFERENCES

Barron, F. Some personality correlates of independence of judgment. *J. Pers.*, 1953, **21**, 287–297.

Bills, R. E., Vance, E. L., & McLean, O. S. An index of adjustment and values. *J. consult. Psychol.*, 1951, **15**, 257–261.

Drevdahl, J. E. Factors of importance for creativity. *J. clin. Psychol.*, 1956, **12**, 21–26.

Flanagan, J. C. The critical incident technique. *Psychol. Bull.*, 1954, **51**, 327–358.

Gough, H. G. *Reference handbook for the Gough Adjective Check List.* Berkeley: Univer. Calif. Inst. of Pers. Assessment and Res., April, 1955.

Stein, M. I. Creativity and culture. *J. Psychol.*, 1953, **36**, 311–322.

Stein, M. I., & Meer, B. Perceptual organization in a study of creativity. *J. Psychol.*, 1954, **37**, 39–43.

Wilson, R. C., Guilford, J. P., & Christensen, P. R. The measurement of individual differences in originality. *Psychol. Bull.*, 1953, **50**, 362–370.

Received July 27, 1958.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 4, 1959

RATIONAL AND MATHEMATICAL RELATIONSHIPS OF SIX SCORING PROCEDURES APPLICABLE TO THREE-CHOICE ITEMS[1]

JACK C. MERWIN

Syracuse University

Multiple-choice items are found on practically all standardized paper-and-pencil tests and on an increasing number of teacher-made examinations. One procedure stands out prominently in the widespread use of multiple-choice items on both published and teacher-made tests: the student is instructed to select what he considers to be the right, or best, answer and is given a score of 1 if he selects the keyed response and a score of 0 if he selects one of the distractors. This procedure is based only on the selection or rejection of the keyed response and has led to suggestions for alternative procedures (Coombs, Millholland, & Womer, 1956; Dressel & Schmid, 1953) which are designed to obtain more information in an attempt to improve the validity of the items.

This paper reports an investigation of the relative validity of three-choice items when each of six procedures is used. Rational and mathematical relationships existing among the six procedures were investigated and item characteristics for which each of the six procedures will yield relatively high and relatively low item validity coefficients are described.

SIX SCORING SCHEMES

The six scoring schemes studied arise from the use of different numbers of item response patterns in combination with different weights. Schemes using two, three, and six response patterns were considered. The weightings used in the schemes were consecutive integral weights and weights which maximize the correlation of the scores with the criterion.

If each subject (S) is instructed to indicate by his response the relative attractiveness, to him, of the alternatives of a three-choice item (using ranking or sequential response procedures), there are just six different response patterns available to him. For example, if the alternatives are "a," "b," and "c," the possible response patterns are "abc," "acb," "bac," "cab," "bca," and "cba." In scoring the item, the tester can utilize all of the information obtained in the S's response and assign one of six scores. He also has the option of disregarding some of the information available in the S's response and assigning one of a lesser number of scores, the number of scores depending on the amount of information he uses. The relationships among the information used in the three scoring patterns investigated in this study are diagrammed in Fig. 1. It is to be recognized that the two-score pattern is the one conventionally used when the Ss are instructed to merely indicate their first choice.

When a six-score scheme is used, all of the information the S gives about the attractiveness of each alternative relative to the others is used. Thus, if "a" is the "best" alternative (or right answer), the person who gives the response pattern "bca" will receive a score different from the person who gives the response pattern "cba," because in the use of the six-score scheme the alternatives are ranked on the key. That is, one of the alternatives must be keyed first, and one of the remaining two must be keyed second. In the example above, the tester must decide whether "c" or "b" is the "second best" alternative. A three-score scheme can be set up in such a

[1] Adapted from the author's doctoral dissertation, *A Mathematical Study of Factors Affecting the Validity of Multiple-Choice Items*, completed at the University of Illinois under the guidance of Lee J. Cronbach.

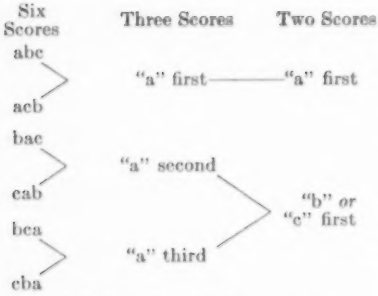

FIG. 1. PATTERNS OF RESPONSE TO BE ASSIGNED SCORES

way that only one alternative need be keyed.

In the three-score scheme studied, the information the *S* reveals about the attractiveness of the nonkeyed alternatives (hereafter called the distractors) relative to one another is not used in assigning scores. Only the information on the attractiveness of the keyed response relative to *each* of the distractors is used. The three scores used in this three-score scheme are assigned on the basis of the rank given to the keyed response. An *S*'s score depends on whether he selects neither, just one, or both distractors ahead of the keyed response. Thus, with this three-score scheme, the response patterns used in the illustration above, "bca" and "cba," receive the same score, since the keyed response, "a," is ranked third in both of them.

In the conventional two-score scheme the concern is only with whether or not the keyed response is the first choice. Here, only information of the attractiveness of the keyed response relative to the most attractive distractor is used. This information can be obtained either by having the *S* make a single selection from among the alternatives or by having him order the alternatives to indicate their attractiveness to him.

Scoring schemes using two, three, or six response patterns could be used with a wide variety of weightings for the different score patterns. For this study, consecutive integral weights were selected because they simplify computation and interpretation; they have generally been found to be satisfactory, and they provide a base of departure to study the gains in efficiency that can be obtained through the use of weights which maximize the correlation between item scores and the criterion.

Guttman (1941) has pointed out that the use of the mean criterion score of those persons who mark each response pattern as the weight for that response pattern maximizes the correlation of the item with the criterion. This leads to an index which is comparable to the correlation ratio, eta. Since this procedure yields the highest possible item-criterion correlation coefficient for any given number of scores, it was deemed appropriate for inclusion in this study of the relative efficiency of scoring schemes. The results for any item, however, will be maximal only for the group used to establish the weights. Some shrinkage in the validity coefficient would be expected under cross-validation.

PROCEDURE

The six scoring schemes studied were compared on the basis of item-criterion correlation coefficients obtained for a variety of items with different parameters.

The validity of an item under any procedure will be determined by the extent to which people high on the criterion get high scores and people low on the criterion get low scores. This information can be expressed in terms of the proportions of groups at different criterion levels who receive each score. In this study, these proportions are the item characteristics which were varied and will be referred to as the *item parameters*.

If the proportions of the criterion groups receiving each score (i.e., item parameters) are known, it is possible to calculate the coefficient of correlation between item scores and the criterion. This coefficient is conventionally considered the "item validity coefficient" and is thought of as a char-

acteristic of the item. In this study, the concern was with the variety of item-criterion correlation coefficients obtained when different scoring procedures are used with the same item and the same criterion. Thus, it was deemed inappropriate to consider the many different coefficients obtained for a given item with the criterion as "item validity coefficients." Any one of the item-criterion correlation coefficients becomes an index of what will be called the *efficiency* of the item under the given scoring procedure. The product-moment correlation coefficient between item scores and the criterion for an item under a given scoring procedure will hereafter be referred to as the *efficiency index* of the item under that scoring procedure.

The relationships reported are based on differences in the efficiency indexes obtained under the various scoring schemes for selected sets of item parameters. It should be recognized that the same difference between indexes at varying positions along the efficiency scale may not represent equal differences in the predictive value of the item. The six procedures were compared on the basis of the resulting efficiency when applied to three-choice items used to predict membership in two or three equally populated criterion groups. Only items which are positively correlated with the criterion were studied.

In a theoretical study of this kind it is not necessary to construct or administer items. The relationships sought were those between item parameters and the resulting efficiency indexes for the various scoring schemes. The relationships were established by systematically varying the item parameters and studying the effects on the efficiency indexes.

Relationships Among Integer Scoring Schemes

The relationships among scoring schemes which utilize two, three, and six consecutive integer scores were studied. The increase in efficiency (as shown by the item-criterion correlation coefficients) obtained by the use of an integer scoring scheme which utilizes more information and yields a larger number of scores than another scheme is directly dependent upon the criterion-relevance of the added information used in assigning scores. If the additional information utilized increases the scores of the high criterion people somewhat more than it increases the scores of the low criterion people, its use in assigning scores will increase the efficiency of the item. The use of additional information to assign a larger number of scores can lead to more, the same, or less efficiency than that obtainable by using less information and a lesser number of scores.

Three Integer Scores vs. Two Integer Scores

The relationship between the efficiency of using three integer scores and the efficiency of using two integer scores can be shown for a variety of item parameters. The item parameters will determine whether the use of three integer scores results in more, the same, or less efficiency than that obtained when the more conventional two-score scheme is used.

Consider the relationships between the two scoring schemes when the two-score scheme has the conventional +1 and 0 scores and the three-score scheme assigns scores of +1, 0, and −1 as the keyed response is ranked first, second, and third, respectively. The selection of these particular consecutive integer scores is for convenience of calculation and comparison; any other set would lead to the same results. As shown in Table 1, in both the two-score and the three-score schemes, a score of +1 is assigned to persons selecting the keyed alternative first. With the two-score scheme, everyone not receiving a score of +1 receives a score of 0. With the three-score scheme, the people who did not choose the keyed response first are split into two groups. One of these groups, those who selected the keyed response second, would receive a score of 0 just as they

TABLE 1

SCORES ASSIGNED UNDER TWO INTEGER SCORING SCHEMES WHEN "a" IS THE KEYED ALTERNATIVE

	Two-Score Scheme	Three-Score Scheme
"a" first	+1	+1
"a" second	0	0
"a" third	0	−1

would under the two-score scheme. The only persons who would receive different scores under the two scoring schemes are those who select both distractors ahead of the keyed alternative. These people receive a score of 0 under the two-score scheme and a score of −1 under the three-score scheme.

Three Criterion Groups. A wide variety of sets of item parameters were used to study the relative efficiency of integer scoring schemes using two and three scores. Whether the three-score scheme will be more efficient, equally efficient, or less efficient than the two-score scheme for a given set of item parameters is indicated by the correlation between the scores of 0 and −1 and the criterion scores for people who do not select the keyed response first. Three examples will serve to indicate the relationship between the magnitude of this correlation coefficient and the relative efficiency of the two scoring schemes.

The efficiency indexes for the two scoring schemes will be the same when there is a small positive correlation between the 0 and −1 scores and the criterion. That is, the mean criterion level of those persons receiving a score of 0 must be slightly higher than the mean criterion level o persons receiving a score of −1 for the two efficiency indexes to be equal. The size of the correlation which is required to make the two indexes equal was found to be in part a function of the mean criterion score of persons who rank the keyed response first and receive a score of +1 under both schemes. Table 2 gives the item parameters for a case in which the efficiency indexes are the same figure, .29. The correlation of the 0 and −1 scores from the three-score scheme with the criterion is .14.

If the correlation of the 0 and −1 scores from the three-score scheme with the criterion is positive and relatively large, the three-score scheme will be superior. Once one of the distractors has been selected first, *S*s are faced with a two-choice situation. They must make a selection between the keyed response and the remaining distractor. If the results of this "two-choice item" are highly enough correlated with the criterion to more than compensate for the difference in scoring under the two procedures, the three-score procedure will be superior. Table 3 presents an illustration of the conditions under which the three-score scheme yields the higher index of efficiency. In this example, the efficiency of the item under the three-score scheme is .81 and under the two-score scheme it is .53. The correlation of the 0 and −1 scores from the three-score scheme with the criterion is .77.

The two-score scheme will be superior when the correlation of the 0 and −1 scores from the three-score scheme with the criterion is negative, zero, or just slightly positive. Table 4 presents an example in which the two-score scheme is superior. In this case the item under the three-score scheme has an efficiency index of .45 and under the two-score scheme the efficiency index is .53. The correlation of the 0 and −1 scores from the three-score scheme with the criterion is .07 for this case.

For the three examples given above, then, correlation coefficients of .14, .77, and .07 between the 0 and −1 scores from the three-score scheme with the criterion resulted in the three-score scheme being equally, more, and less efficient, respectively.

It is possible for the efficiency index of an item under the three-score scheme to surpass the index of the two-score scheme by as much as .41. Such a case is illustrated

TABLE 2

ITEM PARAMETERS FOR WHICH THE EFFICIENCY INDEXES FOR THE TWO-SCORE AND THREE-SCORE SCHEMES ARE EQUAL

(*a*) Proportion of Each Criterion Group Receiving Each Score Under the Three-Score Scheme

	Criterion Group		
	−1	0	1
1	.59	.76	.90
Score 0	.26	.18	.08
−1	.15	.06	.02

Efficiency (r) = .29

(*b*) Proportion of Each Criterion Group Receiving Each Score Under the Two-Score Scheme

	Criterion Group		
	−1	0	1
1	.59	.76	.90
Score 0	.41[a]	.24	.10

Efficiency (r) = .29

[a] Proportion receiving scores of 0 *or* −1 under the three-score scheme.

TABLE 3

ITEM PARAMETERS FOR WHICH THE THREE-SCORE SCHEME IS SUPERIOR TO THE TWO-SCORE SCHEME

(*a*) Proportion of Each Criterion Group Receiving Each Score Under the Three-Score Scheme

	Criterion Group		
	−1	0	1
1	.00	.05	.50
Score 0	.05	.70	.50
−1	.95	.25	.00

Efficiency (r) = .81

(*b*) Proportion of Each Criterion Group Receiving Each Score Under the Two-Score Scheme

	Criterion Group		
	−1	0	1
1	.00	.05	.50
Score 0	1.00[a]	.95	.50

Efficiency (r) = .53

[a] Proportion receiving scores of 0 *or* −1 under the three-score scheme.

TABLE 4

ITEM PARAMETERS FOR WHICH THE TWO-SCORE SCHEME IS SUPERIOR TO THE THREE-SCORE SCHEME

(*a*) Proportion of Each Criterion Group Receiving Each Score Under the Three-Score Scheme

	Criterion Group		
	−1	0	1
1	.00	.05	.50
Score 0	.50	.50	.30
−1	.50	.45	.20

Efficiency (r) = .45

(*b*) Proportion of Each Criterion Group Recieving Each Score Under the Two-Score Scheme

	Criterion Group		
	−1	0	1
1	.00	.05	.50
Score 0	1.00[a]	.95	.50

Efficiency (r) = .53

[a] Proportion receiving scores of 0 *or* −1 under the three-score scheme.

TABLE 5

ITEM PARAMETERS FOR WHICH THE EFFICIENCY INDEX UNDER THE THREE-SCORE SCHEME IS .41 GREATER THAN THE INDEX UNDER THE TWO-SCORE SCHEME

(*a*) Proportions of Each Criterion Group Receiving Each Score Under the Three-Score Scheme

	Criterion Group		
	−1	0	1
1	.00	.33	.33
Score 0	.00	.67	.67
−1	1.00	.00	.00
Efficiency (*r*) = .74			

(*b*) Proportions of Each Criterion Group Receiving Each Score Under the Two-Score Scheme

	Criterion Group		
	−1	0	1
1	.00	.33	.33
Score 0	1.00[a]	.67	.67
Efficiency (*r*) = .33			

[a] Proportion receiving scores of 0 or −1 under the three-score scheme.

in Table 5. In this case, the efficiency index of the item under the three-score scheme is .74, and the index under the two-score scheme is .33.

Two Criterion Groups. An evenly divided dichotomous criterion was used to seek further generalizations about the relative efficiency of items under the two-score and three-score schemes. The main finding of this part of the study was that the three-score scheme will more often than not lead to greater efficiency than the two-score scheme. More specifically, when two equal sized criterion groups are considered, the following can be concluded: (*a*) If the probability that persons low on the criterion will rank one of the distractors first is equal to the probability that they will rank the other one first, the three-score scheme will be more efficient than the two-score scheme for nearly all possible responses from persons high on the criterion. (*b*) If persons low on the criterion who do not rank the keyed response first have a large probability of ranking it second, the three-score scheme will be more efficient than the two-score scheme only when persons high on the criterion who do not rank the keyed response first also have a large probability of ranking it second. (*c*) If persons low on the criterion have a very low probability of choosing the keyed response first, the three-score scheme will be more efficient unless persons high on the criterion have a very high or very low probability of choosing the keyed response first.

In any comparison of product-moment correlation coefficients as measures of efficiency for the two scoring schemes, the dependency of the maximum possible coefficient on the number of criterion groups must be considered. The two-score scheme cannot yield a coefficient of 1.00 when more than two criterion groups are utilized and the three-score scheme cannot yield an index of perfect efficiency when more than three criterion groups are used. For three equally populated criterion groups, the maximum coefficient for the three-score scheme is, of course, 1.00; for the two-score scheme it is .866. When considering the relative efficiency of the two scoring schemes with three equal sized criterion groups, this limit on the efficiency index should be considered. When two criterion groups are used, the maximum possible efficiency index is 1.00 for both schemes.

Six Integer Scores vs. Three Integer Scores

The use of six consecutive integer scores involves the use of all the information utilized in the three-score scheme plus the information on the relative attractiveness

of the distractors to each other. With "a" keyed as the best answer and "b" keyed as the "second best" answer, the response patterns "abc," "acb," "bac," "cab," "bca," and "cba" might be assigned scores of 5, 4, 3, 2, 1, and 0, respectively. Such an assignment of scores is arbitrary, and any set of consecutive integers would lead to the same results.

Whether the six-score scheme will lead to an index of efficiency which is larger, the same as, or smaller than that obtainable through the use of the three-score scheme will be determined by the relationship of the additional information utilized in assigning scores to the criterion. Thus, if alternative "b" is keyed "second best" and alternative "c" is keyed "third best," the extent to which high criterion persons choose "b" ahead of "c" and low criterion persons choose "c" ahead of "b" will determine the possible increase in efficiency of the six-score scheme over the three-score scheme. A single, though atypical, example which is heavily loaded in favor of the six-score scheme can serve to point up the fact that the gains in efficiency possible through the use of six integer scores are limited.

Consider two criterion groups for which the *only* criterion-relevant information is to be found in the relative ranking of the two distractors (Table 6). Both groups rank the keyed response first, second, and third with equal frequency, the low criterion group always ranking "c" ahead of "b" and the high criterion group always ranking "b" ahead of "c". For this set of item parameters the predictive efficiency of the two-score and three-score schemes would be zero. For the six-score scheme the index of efficiency would be .29. Thus, even with item parameters which are extremely favorable to the six-score scheme, and highly improbable in practice, the gain in efficiency is only .29.

In general, a study of a variety of sets of item parameters indicates that relatively small gains in efficiency can be expected through the use of six integer scores rather than three integer scores.

TABLE 6

ITEM PARAMETERS FOR WHICH THE EFFICIENCY INDEXES FOR THE SIX-SCORE, THREE-SCORE, AND TWO-SCORE SCHEMES ARE .29, .00, AND .00, RESPECTIVELY

		Criterion Group	
		0	1
Response Pattern	abc	.00	.33
	acb	.33	.00
	bac	.00	.34
	cab	.34	.00
	bca	.00	.33
	cba	.33	.00

CONTRIBUTION OF MAXIMIZING WEIGHTINGS

The procedure given by Guttman for determining weights which will maximize the item-criterion correlation coefficient was discussed earlier. We will now consider the contribution of such weights to the efficiency of the item.

The Two-Score Scheme

The product-moment correlation coefficient, used in this study as the index of the efficiency of an item under a given scoring scheme, is not affected by the size of the scores when just two scores are used. Thus, the simple use of integer scores is just as efficient, but no more efficient, than any other weighting that might be used for the two scores.

The Three-Score Scheme

The use of weights which maximize the correlation of item scores with the criterion adds most to the efficiency of the item when the relationship between the item responses and the criterion deviate from linearity as much as possible. These are also the conditions under which the three-score integer scheme has the greatest superiority over the two-score integer scheme.

We can examine a case where maximizing weights make an extreme contribution to the efficiency of the item.

Consider three equal sized criterion groups with the item parameters given in Table 5. The low criterion group always ranks the keyed response third. The middle and top groups both rank the keyed response first with a probability of .33 and second with a probability of .67, never ranking it third. Under these conditions, the efficiency index for the item for the three-score scheme with maximizing weights is .87. If three integer scores were used, the index would be .74, and if two scores were used, it would be .33. Thus, under conditions most favorable to the efficiency of maximizing weights, the increase in efficiency of the maximizing weights over integer weights is only .13, and the increase in efficiency of three integer scores over two integer scores is .41.

A study of a variety of sets of item parameters indicates that the use of maximizing weights can lead to only relatively small increases in the efficiency of an item over that obtained when three integer scores are used.

The Six-Score Scheme

The use of a six-score scheme with maximizing weights for the responses represents the maximum use of all the information that is obtained when subjects rank the responses of a three-choice item. While other schemes may yield indexes (for specifiable sets of item parameters) as large as that obtained by this scheme, under no conditions can they be larger.

The set of item parameters under which the use of six maximally weighted scores will produce the greatest gain in efficiency over other procedures is that set presented earlier (Table 6) to compare the efficiency of six integer scores to the efficiency of three integer scores. In this case, all of the information about two criterion groups is in the ranking of the two distractors, the low criterion group always ranking "c" ahead of "b," and the high criterion group always ranking "b" ahead of "c." As was pointed out earlier, this set of item parameters leads to an efficiency index of .00 for both a two-score and a three-score scheme, and this would be true regardless of the weightings used. Six integer scores would yield a validity coefficient of .29, and six scores maximally weighted would give a coefficient of 1.00. With these particular item parameters the advantage of this procedure over the others is substantial, but this particular set of parameters is not to be expected in practice. The only conditions under which the use of maximizing weights can lead to an efficiency index of 1.00 is when each response pattern is given by persons from no more than one criterion group.

A study of a number of different sets of item parameters indicates that for sets of item parameters to be expected in practice the efficiency of an item under the other procedures will not be much less than that obtained through the use of six maximally weighted scores.

Summary

In the conventional multiple-choice test procedure, a subject receives a score of +1 if he chooses the keyed alternative, otherwise he receives a score of 0. This study compares the relative efficiency of three-choice items with specified characteristics when used with six different scoring procedures. The product-moment correlation coefficient between item scores and a criterion was used as the index of efficiency.

Scoring schemes which utilize two, three, and six consecutive integer scores and sets of two, three, and six scores weighted to produce maximum correlation with the criterion were compared. They were compared on the bases of (*a*) the information used in assigning scores and (*b*) the efficiency with which they can be used to predict membership in two or three equally populated criterion groups.

The following conclusions were drawn from the study:

1. The use of added information to assign

an increased number of integer scores does not necessarily lead to greater efficiency of an item. It can lead to greater, identical, or lesser efficiency than that obtainable by using a smaller number of integer scores.

2. If added information is used to assign more integer scores, the resulting efficiency will be a function of the criterion-relevance of the additional information utilized.

3. The use of three integer scores will more often than not lead to greater efficiency than can be obtained with two integer scores when two equal-sized criterion groups are used.

4. The use of six integer scores rather than three integer scores will not, in general, lead to a substantial increase, and can lead to a decrease, in efficiency.

5. The use of six maximally weighted scores will always lead to item efficiency which is as high, or higher than that obtainable by any other procedure. However, differences between the index from this procedure and indexes from other procedures are relatively small for sets of item parameters to be expected in practice. Shrinkage of cross-validation would be expected to make these differences even smaller.

REFERENCES

COOMBS, C. H., MILLHOLLAND, J. E., & WOMER, F. G. The assessment of partial knowledge. *Educ. psychol. Measmt.*, 1956, **16**, 13-37.

DRESSEL, P. L., & SCHMID, J. Some modifications of the multiple-choice item. *Educ. psychol. Measmt.*, 1953, **13**, 574-595.

GUTTMAN, L. Mathematical and tabulation techniques. In P. Horst (Ed.), *The prediction of personal adjustment*. New York: Social Science Research Council, 1941. Pp. 251-346.

Received August 5, 1958.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 4, 1959

INTERPRETATION OF DIFFERENCES AMONG AVERAGES AND INDIVIDUAL TEST SCORES

FREDERICK B. DAVIS
Hunter College

The interpretation of test scores frequently involves differences among scores, among averages, and among scores and averages. For example, a school psychologist or counselor may wish to compare scores on the same test that were obtained by two different pupils, to compare the mean scores on the same test obtained by two particular classes, or to compare a pupil's score with the mean score of his class on the same test. Such comparisons involve differences among scores, differences that can be interpreted scientifically only if their standard errors of measurement are known. As long ago as 1923, the standard error of measurement of a difference between two individual standard scores was presented (Kelley, 1923); in 1928, the standard error of measurement of the mean was provided (Huffnaker & Douglass, 1928). But, to the best of the writer's knowledge, the standard errors of measurement of several useful differences between individual and mean scores have not been published. The main purpose of this paper is to present some of these, together with their applications and derivations. They make possible estimation of the probability that a given difference is a chance deviation from a true difference of zero.

In practice, differences between test scores large enough to have a probability of occurring by chance 15 times, or less, out of 100 may be regarded as worth interpreting. This level of significance for the two-tailed test of the null hypothesis may seem unduly lenient to psychologists accustomed to using the .01 or the .05 levels in experimental work. It must be remembered, however, that the designation of any level of probability as "significant" is arbitrary and represents a balance between the test interpreter's desire to avoid accepting differences as attributable to something other than chance when in fact they are not and his desire to avoid attributing differences to chance when in fact they are not. For interpreting test scores, several factors suggest a rather lenient level of significance as appropriate. First, scores derived from most achievement and aptitude tests are sufficiently unreliable as to make their practical utility doubtful if only differences among individual scores significant at a stringent level (such as .01) are interpreted. Second, the penalty for accepting a difference as owing to something other than chance when, in fact, it is a chance deviation from a true difference of zero is not usually great, because test results are ordinarily only one of several factors entering into the making of any important decision about a child's schooling. Practical experience suggests the .15 level as appropriate for interpreting differences among the test scores of school children and it is, therefore, most often used in this paper. However, the basic equations permit choice of any level of significance that is desired.

For clarity of presentation and for the convenience of test users, eight common types of differences among scores have been identified and treated separately with illustrative examples.

Case 1: A Difference Between the Average of Several Scores Obtained by One Individual and His Score on One of the Tests Included in the Average

Clinical and school psychologists often wish to compare the average of several test scores obtained by one individual with his score on one of the tests included in the average. For example, an individual's average score on 10 tests in the Wechsler Adult Intelligence Scale (excluding the Vocabu-

TABLE 1

PART SCORES OF 18-YEAR-OLD BOY ON WAIS, AND DATA PERTAINING TO STANDARD ERRORS OF MEASUREMENT OF DIFFERENCES BETWEEN AVERAGE AND PART SCORES[a]

Part	Boy's Scaled Score	Difference: Deviation from Average	Standard Error of Measurement of Difference[b]	Minimum Deviation From Average Significant at 5% Level
1. Information	10	+0.5	.89	±1.74
2. Comprehension	11	+1.5	1.29	±2.53
3. Arithmetic	9	−0.5	1.30	±2.55
4. Similarities	13	+3.5	1.08	±2.12
5. Digit Span	7	−2.5	1.52	±2.98
7. Digit Symbol	9	−0.5	.87	±1.71
8. Picture Completion	9	−0.5	1.14	±2.23
9. Block Design	11	+1.5	1.12	±2.20
10. Picture Arrangement	9	−0.5	1.59	±3.12
11. Object Assembly	7	−2.5	1.54	±3.02
Sum of scaled scores	95			
Average scaled score	9.5			
6. Vocabulary	11	+1.5	.81	±1.59

[a] The standard errors of measurement of differences and the minimum deviations from an individual's average scaled score that are significant at the .05 level are generally applicable and may be used to interpret the WAIS scores of any 18- to 19-year-old. As a matter of fact, they are serviceable in the interpretation of scores of individuals at any age for which the test is appropriate.

[b] Based on standard errors of measurement given in the Manual (Wechsler, 1955, p. 13).

lary test) may be compared with his score on each one of the tests included in the average.[1] Table 1 shows the scaled scores obtained by an 18-year-old boy on the WAIS. Whether the difference between this boy's average scaled score of 9.5 and any one of his scaled scores on the separate tests is statistically significant can be determined at any desired level by comparing each difference with its own standard error of measurement. The required standard errors of measurement are presented in Table 1, together with the minimum deviation of each part score from the average that can be regarded as significant at the .05 level. The data indicate that only the boy's scaled score of 13 on the Similarities test deviates sufficiently from his average scaled score on 10 tests to warrant the inference that the deviation is attributable to something other than chance. It should be noted that if the 10 differences in a given individual's set of WAIS scores were independently determined by chance, one difference in every two sets would be expected to show significance at the .05 level. When at least two differences in a given set are found to be significant at that level, it is unlikely that their occurrence can be attributed to chance, even though there is not complete independence in their determination.

The standard error of measurement of the difference between an average score obtained by any individual chosen at random from the group tested and any one of the scores that entered into his average may

[1] Wechsler (1944, Ch. 11) has discussed comparisons of this sort. More recently, he has discussed comparisons of the sort covered by Case 2 in this paper (Wechsler, 1958, Ch. 11).

be estimated by Equation [1]:

$$s_{meas_{((T/m)-Z_I)}} = \sqrt{\frac{s^2_{meas_T}}{m^2} + \left(\frac{m-2}{m}\right) s^2_{meas_{Z_I}}} \quad [1]$$

where $s^2_{meas_T}$ is the variance error of measurement of the sum of the m part scores, m is the number of parts for which the scores are comparable, T/m is the average of the comparable part scores, and $s^2_{meas_{Z_I}}$ is the variance error of measurement of any one of the comparable part scores. Sometimes it is convenient to obtain $s^2_{meas_T}$ by summing the variance errors of measurement of the part scores, thus making use of the fact that

$$s^2_{meas_T} = \sum^{m} s^2_{meas_{Z_I}}$$

Whether the difference between an individual's average of m comparable scores and his score on one of the m parts is significantly different from zero at any desired level may be estimated with serviceable accuracy when the standard error of measurement was computed in a large sample by means of the critical ratio:

$$CR = \frac{((T/m) - Z_I) - 0}{s_{meas_{((T/m)-Z_I)}}} \quad [2]$$

Differences between an average and scores on any one of the tests included in it are convenient to interpret only if the separate tests yield comparable scores. The latter are defined, for purposes of this paper, as transformed obtained scores for which the corresponding points in the distributions of true scores are exceeded by the same percentages of examinees in a defined sample. The transformation is made simply to cause the comparable scores to be numerically identical. It is not necessary that comparable scores be measures of the same ability or that they be equally reliable. Serviceable approximations to comparable scores, as so defined, may be obtained by the method described by Flanagan (1951). They have been provided, though by a different method, for scores on the parts of the WAIS.

Derivation of Equation [1]. Let T represent an individual's total score made up of the sum of m part scores, any one of which may be denoted as I. Then

$$T = A + B + \cdots + M$$

and

$$\frac{T}{m} = \frac{\sum^{m} I}{m}$$

the average of the part scores. Suppose that an indefinitely large number (n) of equivalent forms of the m parts were administered to one pupil chosen at random from a group for which the tests are appropriate, and postulate that the pupil's true score in each of the abilities measured remained constant throughout the testing. A distribution of the n differences between the pupil's n average scores and his n scores on any given part would be produced. These differences would be normally distributed around a mean approaching $(T_t/m) - I_t$, the difference between the pupil's true mean and true score. Since it was postulated that the pupil's true scores would remain constant, $\sigma^2_{T_{t/m}} = \sigma^2_{I_t} = 0$. Consequently, the population variance of the distribution of differences may be written as:

$$\sigma^2_{meas_{((T/m)-I)}} = \sigma^2_{T_{e/m}} + \sigma^2_{I_e} - 2\sigma_{T_{e/m}}\sigma_{I_e}\rho_{(T_{e/m})I_e} \quad [3]$$

where the subscript e denotes error of measurement, σ^2 denotes a population variance, and ρ denotes the product-moment correlation coefficient in the population.

It can easily be shown that

$$\rho_{(T_e/m)I_e} = \frac{\sigma_{I_e}}{\sigma_{T_e}} \quad [4]$$

Hence, Equation [3] may be simplified to:

$$\sigma^2_{meas_{((T/m)-I)}} = \frac{\sigma^2_{T_e}}{m^2} + \left(\frac{m-2}{m}\right)\sigma^2_{I_e} \quad [5]$$

In practice, if the variance errors of meas-

urement have been computed on the basis of a large sample, they may be used in place of the population variance errors of measurement in Equation [5]. If scores on all parts of total score T in Equation [5] are comparable, we may denote them with the capital letter Z and write:

$$s^2_{T_e} = s^2_{meas_T}$$
$$= s^2_{(Z_{Ae}+Z_{Be}+\cdots+Z_{Me})}$$
$$= s^2_{Z_{Ae}} + s^2_{Z_{Be}} + \cdots + s^2_{Z_{Me}}$$
$$= \sum^{m} s^2_{meas_{Z_I}}$$

since the intercorrelations of the errors of measurement in separate tests approach zero for an indefinitely large number of pairs of scores administered under proper conditions.

By definition, $s^2_{meas_{Z_I}} = s^2_{Z_I}(1 - r_{Z_I Z_{I'}})$. The transformation of raw scores to comparable scores, as defined above, has no appreciable effect on their correlations with other variables; hence, $r_{II'} \equiv r_{Z_I Z_{I'}}$ if we assume that transformation to have no effect on the reliability coefficient of part-score I of total score T. Then:

$$s^2_{meas_{Z_I}} = s^2_{Z_I}(1 - r_{II'})$$
$$= \left(\frac{s^2_{Z_I}}{s^2_I}\right) s^2_I (1 - r_{II'}) = \frac{s^2_{Z_I} s^2_{meas_I}}{s^2_I}$$

By making appropriate substitutions of these values in Equation [5], Equation [1] is easily obtained.

Case 2: A Difference Between the Average of Several Scores Obtained by One Individual and His Score on a Test Not *Included in the Average*

The Vocabulary scaled score of the boy whose WAIS scores are presented in Table 1 was excluded from the computation of his average scaled score of 9.5. Hence, the product-moment correlation of errors of measurement in the Vocabulary score and in his average scaled score would be insignificantly different from zero if an indefinitely large number of equivalent forms of the test were administered to him. If we assume this correlation to be zero, as is customary, the analogue of Equation [3] for the population variance error of measurement of this difference between the average of several scores obtained by an individual and his score on a test *not* included in the average may be written as:

$$\sigma^2_{meas_{((T/m)-V)}} = \sigma^2_{T_{e/m}} + \sigma^2_{V_e}$$

This leads to Equation [6], which may be used to estimate the standard error of measurement of a difference of this kind when a large sample is used in computing the variance errors of measurement:

$$s_{meas_{((T/m)-ZV)}} = \sqrt{\frac{s^2_{meas_T}}{m^2} + s^2_{meas_{Z_V}}} \qquad [6]$$

The value yielded by Equation [6] may be used in a critical ratio analogous to Equation [2] for determining the statistical significance of a difference between the average of several scores obtained by one individual and his score on a test *not* included in the average. It should be recalled that such differences are inconvenient to interpret unless all of the tests yield comparable scores.

Case 3: A Difference Between the Average Score of Several Specified Individuals on One Test and the Score on the Same Test of One of the Individuals Included in the Average

At the end of the school year, the 10 pupils in Miss Black's small fifth-grade class obtained the following grade scores on the Paragraph Meaning Test of the Stanford Achievement Test, Intermediate Form J: 97, 81, 72, 64, 60, 58, 51, 47, 43, 35. The average grade score was 60.80, and the standard deviation of scores was 17.70. The standard error of measurement of the difference between any one of these 10 scores, chosen at random, and their average may be estimated by Equation [7].

$$s_{meas_{(\bar{X}-X)}} = s_{meas_X}\sqrt{\frac{N-1}{N}} \qquad [7]$$

where N is the number of pupils whose scores were averaged, and a large sample was used in computing the standard error of measurement.

The standard error of measurement reported by the publisher of the Stanford Paragraph Meaning Test is 6 grade-score points in a sample of 243 fifth-grade pupils in which the mean grade score was about 58 and the standard deviation was about 19 grade-score points. For the difference between the mean grade score of the 10 pupils and the score of any one pupil in Miss Black's class, Equation [7] yields a standard error of measurement of 5.70. To determine the smallest difference significant at the .15 level, 5.70 may be multiplied by 1.44, which is the approximate critical ratio when the standard error of measurement has been computed in a sample as large as 243. The result (rounded to the nearest tenth) is 8.2. Thus, only scores of 97, 81, 72, 51, 47, 43, and 35 are found to be significantly different at the .15 level from the average of 60.80.

Derivation of Equation [7]. Let X_{1A}, X_{1B}, $\cdots$, X_{1N} represent the raw scores of each of N pupils on Form 1 of Test X, and let $\overline{X}_1$ represent the mean of these scores. Assume that an indefinitely large number (n) of equivalent forms of the test are administered to the same N pupils, and postulate that the true score of each pupil remains constant throughout the testing. Then X_{2A}, X_{3A}, $\cdots$, X_{nA} represent additional raw scores of pupil A, etc.; and $\overline{X}_2$, $\overline{X}_3$, $\cdots$, $\overline{X}_n$ represent additional mean scores of the N pupils. For any pupil, say Pupil A, an essentially normal distribution of differences will be produced between the mean of the n means and the pupil's scores on successive forms of the test. The mean of this distribution of differences will approach $\overline{X}_t - X_{tA}$, where $\overline{X}_t$ is the true mean of the N pupils and X_{tA} is the true score of Pupil A. Since it was postulated that the true score of each pupil would remain constant throughout the testing, the variance of the distribution of differences is generated only by errors of measurement and may be written as:

$$\sigma^2_{(\overline{X}-X_A)} = \sigma^2_{\overline{X}_e} + \sigma^2_{X_{eA}} - 2\sigma_{\overline{X}_e}\sigma_{X_{eA}}\rho_{\overline{X}_e X_{eA}} \quad [8]$$

where $\sigma^2_{\overline{X}_e}$ is the population variance error of measurement of the mean, $\sigma^2_{X_{eA}}$ is the population variance error of measurement of Pupil A's score, and $\rho_{\overline{X}_e X_{eA}}$ is the correlation of the errors of measurement in the n successive means and the n successive raw scores of Pupil A. This correlation may be expressed as:

$$\rho_{\overline{X}_e X_{eA}} = \frac{1}{\sqrt{N}} \quad [9]$$

The variance error of measurement of the mean may be written as (Huffnaker & Douglass, 1928):

$$\sigma^2_{\overline{X}_e} = \frac{\sigma^2_X(1 - \rho_{XX'})}{N} = \sigma^2_{meas_{\overline{X}}} \quad [10]$$

It is well known that for any pupil drawn at random from an indefinitely large sample,

$$\sigma^2_{X_{eA}} = \sigma^2_{meas_X} = \sigma^2_X(1 - \rho_{XX'}) \quad [11]$$

If Equations [9], [10], and [11] are substituted in Equation [8], the latter reduces to:

$$\sigma^2_{(\overline{X}-X_A)} = \left(\frac{N-1}{N}\right)\sigma^2_{meas_X} \quad [12]$$

where N is the number of pupils whose scores were averaged.

The population variance error of measurement may be estimated as

$$s^2_{meas_X} = N_P s^2_{meas_X}/(N_P - 1)$$

where N_P is the number of pupils in the sample in which $s^2_{meas_X}$ was computed. It is interesting to note that if $N_P = N$, Equation [12] becomes

$$s^2_{meas_{(\overline{X}-X_A)}} = s^2_{meas_X}$$

In practice, if $s^2_{meas_X}$ has been computed on the basis of a large sample, it may be used in place of $\sigma^2_{meas_X}$ in Equation [12] to obtain Equation [7].

Case 4: A Difference Between the Average Score of Several Specified Individuals on One Test and the Score of an Individual Not *Included in the Average*

A fifth-grade pupil *not* in Miss Black's class was tested at about the same time as the members of her class with the same test. His grade score was 69. The standard error of measurement of the difference between his score of 69 and the mean score of 60.80 in Miss Black's class may be estimated by Equation [13]:

$$s_{meas_{(\bar{X}-X_0)}} = s_{meas_X}\sqrt{\frac{N+1}{N}} \qquad [13]$$

Since the standard error of measurement of the Paragraph Meaning Test is 6 grade-score points, Equation [13] yields a standard error of measurement of the difference of 6.3 between the average score of 60.80 and the score of the pupil *not* included in the average. The smallest difference significant at the .15 level is, therefore, 9.1 points (rounded to the nearest tenth), and we conclude that the pupil's grade score of 69 is not significantly different from the average score in Miss Black's class of 10 pupils.

Equation [13] is derived in the same manner as Equation [7], but the analogue of Equation [8] in its derivation includes a covariance term that equals zero. Therefore, Equations [7] and [13] differ.

Case 5: A Difference Between the Average Score of a Representative Sample of a Defined Population and the Score on the Same Test of a Particular Individual Not *Included in the Sample*

A twelfth-grade pupil, tested early in November, obtained a scaled score of 68 in Speed of Comprehension on the Davis Reading Test, Series 1. To determine whether his score differs significantly from the average of the norms group of twelfth-grade pupils, Equation [14] may be used:

$$s_{(\bar{X}-X_P)} = \sqrt{\frac{s_X^2}{N-1} + \left(\frac{N_P}{N_P-1}\right)s_{meas_X}^2} \qquad [14]$$

where s_X^2 and N are, respectively, the variance and number of pupils in a representative sample from a defined population, X_P is a particular individual *not* included in this sample, and N_P is the number of pupils in the sample used to compute $s_{meas_X}^2$.

The norms group of twelfth-grade pupils tested in September and October of 1957 included 5596 boys and girls in which the mean and standard deviation of Speed of Comprehension scores were 71.7 and 7.1, respectively. The standard error of measurement is given in the test manual as 2.9 in a sample of 1096 twelfth-grade pupils in which the mean and standard deviation were, respectively, 73.5 and 7.3. For these data, Equation [14] yields a value of 2.90. Since the smallest deviation from the average score of 71.7 in the norms group that is significant at the .15 level is, therefore, 4.2 (rounded to the nearest tenth), we cannot conclude that the pupil's score of 68 differs significantly from the average score of pupils at his grade level.

It is obvious that Equation [14] has general utility for establishing the significance of a difference between the mean score of a sample drawn at random from a defined population and the score of a particular individual. It may be noted that when the sample is very large, the value yielded by Equation [14] will be essentially the same as the standard error of measurement of the individual's obtained score. The variance of means in Equation [14] is that of the means of successive samples drawn at random from a defined population, whereas in Equation [13] it is that of the means of the same sample of individuals on successive equivalent forms of the same test.

Case 6: A Difference Between the Mean Scores of One Individual on n *Equivalent Forms of Test X and on* m *Equivalent Forms of Test Y, When Scores on Tests X and Y Are Comparable*

To increase accuracy of measurement, more than one of the equivalent forms of a test may be administered to one indi-

vidual. It is well known that the standard error of measurement of the mean score on n such forms of, say, Test X is:

$$s_{meas\bar{X}} = \frac{s_{meas_X}}{\sqrt{n}} \qquad [15]$$

It can easily be shown that the standard error of measurement of the difference between the mean of n equivalent forms of Test X administered to one individual and the mean of m equivalent forms of Test Y administered to the same individual or to another individual can be estimated by Equation [16]:

$$s_{meas(\bar{X}-\bar{Y})} = \sqrt{\frac{s^2_{meas_X}}{n} + \frac{s^2_{meas_Y}}{m}} \qquad [16]$$

If the numbers of cases on which the variance errors of measurement in Equation [16] are large, the value yielded by this equation may be used in a critical ratio analogous to Equation [2] for estimating the statistical significance of the difference between two means of equivalent forms.

Case 7: A Difference Between the Mean Scores on the Same Test of Two Particular Partially Overlapping Groups

Miss Black's fifth-grade class of 10 pupils is one of 14 fifth-grade classes in Smithville. The mean grade score of the 245 pupils in these classes was 64.00 when they were tested with the Stanford Paragraph Meaning Test on the same date. Whether the difference between the mean score of these particular 245 pupils in Smithville and the mean score of 60.80 in Miss Black's class is significant can be determined by comparing the difference of 3.20 grade-score points with the estimated standard error of measurement of the difference between the means of these partially overlapping groups. This standard error of measurement may be estimated by Equation [17]:

$$\hat{s}_{meas_{(\bar{X}_T-\bar{X}_S)}} = s_{meas_X} \sqrt{\frac{N_P}{N_P - 1}\left(\frac{1}{N_S} - \frac{1}{N_T}\right)} \qquad [17]$$

where $\bar{X}_T$ and $\bar{X}_S$ are the means and N_T and N_S the numbers of cases in the total group and subgroup, respectively; and s_{meas_X} is the standard error of measurement computed on the basis of N_P cases having a mean and standard deviation similar to those in the total group.

The standard error of measurement for the Stanford Paragraph Meaning Test is given by the publisher as 6 grade-score points in a sample of 243 fifth-grade pupils. Equation [17] yields a value of 1.84, and the t ratio is 1.74, with 242 degrees of freedom. Hence, we conclude that the difference between the mean of Miss Black's class and the mean of all pupils in the 14 classes (including Miss Black's) is significant at better than the .10 level.

Derivation of Equation [17]. Suppose that an indefinitely large number (n) of equivalent forms of Test X were administered to every individual in Group T, and postulate that the true score of each individual remained constant throughout the testing. If the mean score of Group T were computed for each of the n equivalent forms, n means would be obtained. If the mean score of Subgroup S of Total Group T were computed for each of the n equivalent forms, n means for Subgroup S would likewise be obtained. An essentially normal distribution of differences beween the two means for each form would be produced. The mean of these differences would approach $\bar{X}_{tT} - \bar{X}_{tS}$, where the subscript tT denotes a true score in Group T and the subscript tS a true score in Subgroup S. The variance of these differences may be expressed as:

$$\sigma^2_{meas_{(\bar{X}_T-\bar{X}_S)}} = \sigma^2_{(\bar{X}_{tT}+\bar{X}_{eT}-\bar{X}_{tS}-\bar{X}_{eS})}$$

where the subscript eT denotes an error of measurement in Group T and eS an error of measurement in Subgroup S. In terms of individual scores, this may be written as:

$$\sigma^2_{meas_{(\bar{X}_T-\bar{X}_S)}} = \sigma^2\left(\frac{\sum^{N_T} X_{tT}}{N_T} + \frac{\sum^{N_T} X_{eT}}{N_T} - \frac{\sum^{N_S} X_{tS}}{N_S} - \frac{\sum^{N_S} X_{eS}}{N_S}\right)$$

which will be found equal to:

$$\sigma^2_{meas_{(\bar{X}_T - \bar{X}_S)}} = \frac{\sum^{N_T} \sigma^2_{eT}}{N^2_T} + \frac{\sum^{N_S} \sigma^2_{eS}}{N^2_S} - \frac{2\sum^{N_S} \sigma_{eT}\sigma_{eS}}{N_T N_S} \quad [18]$$

If we assume that σ^2_{eT} is the same for every individual in Group T and that $\sigma^2_{eT} = \sigma^2_{eS}$, Equation [18] may be simplified to:

$$\sigma^2_{meas_{(\bar{X}_T - \bar{X}_S)}} = \sigma^2_{eT}\left(\frac{1}{N_S} - \frac{1}{N_T}\right)$$

from which Equation [17] may be obtained.

Case 8: A Difference Between the Mean Scores of Two Particular Nonoverlapping Groups on Any Variable

The pupils in Miss Green's class and Miss Brown's class were given an algebra test on approximately the same date at the end of the school year. The tests were carefully administered by the same examiner, whose records indicate that each pupil cooperated well. The results are summarized in Table 2. Since the variances of obtained scores in the two classes are not significantly different ($F = 1.29$), we assume equivalent variances of the scores in the population or populations from which Miss Green's and Miss Brown's classes were drawn. If we also assume these scores to be distributed normally in the population or populations, we may test the difference between the sample means to see whether it is significantly different from zero. The t ratio is:

$$t = \frac{(\bar{X}_G - \bar{X}_B) - 0}{\hat{s}_{X_T}\sqrt{\frac{1}{N_G} + \frac{1}{N_B}}}$$

where:

$$\hat{s}_{X_T} = \sqrt{\frac{\Sigma(X_G - \bar{X}_G)^2 + \Sigma(X_B - \bar{X}_B)^2}{N_G + N_B - 2}}$$

Since $t = .80$, with 62 degrees of freedom, we conclude that the two classes may well have been drawn at random from the same parent population. This result indicates that differences between the means of pairs of classes of the sizes of Miss Green's and Miss Brown's drawn successively at random from the population of such classes could be expected by chance often to be as large or larger than the difference of 2.18 points between the means of Miss Green's and Miss Brown's classes. But this result does not tell us whether a difference as large as 2.18 points could often occur by chance between the means of these two classes if they were tested over and over with equivalent forms of the test and if the true difference were zero. Yet this may be exactly what we want to know. To determine this, a different t ratio should be employed, with the estimated standard error of measurement of the difference between the means as the denominator. It can easily be shown that the standard error of measurement of a difference between the means of two particular nonoverlapping groups on any variable, say Test X, may be estimated as:

$$\hat{s}_{meas_{(\bar{X}_G - \bar{X}_B)}} = s_{meas_X}\sqrt{\frac{N_P}{N_P - 1}\left(\frac{1}{N_G} + \frac{1}{N_B}\right)} \quad [19]$$

where s_{meas_X} was computed in a sample of N_P cases and N_G and N_B are the numbers of cases used in computing $\bar{X}_G$ and $\bar{X}_B$, respectively.

The standard error of measurement of the algebra test given to Miss Green's and

TABLE 2
SCORES IN TWO CLASSES

	Miss Green's Class	Miss Brown's Class
N	34	30
$\bar{X}$	28.68	26.50
s_X	11.28	9.95
$s_{\bar{X}}$	1.97	1.85
$r_{XX'}$	.89	.86
s_{meas_X}	3.72	3.72

Miss Brown's classes was computed on the basis of the 64 pupils in both classes, the mean errors of measurement being insignificantly different in the two classes. For the difference between the means of the two algebra classes, we have $t = 2.32$, with 62 degrees of freedom, and we conclude that the mean ability in algebra of these two particular classes is almost surely different, Miss Green's class being superior.

As is well known, care must be exercised in choosing the appropriate denominator for the t ratio. In making this choice, the experimenter should ask this question: Do I want to find whether the algebraic signs would rather consistently be the same for differences between means of scores in successive samples drawn at random from the same population, or for differences between means of scores on successive equivalent forms administered to the same sample? If the former is desired, an appropriate standard error of the difference between means should be employed; if the latter is desired, an appropriate standard error *of measurement* of the difference between means should be employed.

REFERENCES

FLANAGAN, J. C. Units, scores, and norms. In *Educational measurement*. Washington: American Council on Education, 1951. Pp. 752–760.

KELLEY, T. L. A new method for determining the significance of differences in intelligence and achievement scores. *J. educ. Psychol.*, 1923, **14**, 321–333.

HUFFNAKER, C. L., & DOUGLASS, H. R. On the standard errors of the mean due to sampling and to measurement. *J. educ. Psychol.*, 1928, **19**, 643–649.

WECHSLER, D. *Measurement of adult intelligence*. (3rd ed.). Baltimore: Williams & Wilkins, 1944.

WECHSLER, D. *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corp., 1955.

WECHSLER, D. *Measurement and appraisal of adult intelligence*. (4th ed.). Baltimore: Williams & Wilkins, 1958.

Received December 3, 1958.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 4, 1959

EFFECTS OF "BRAINSTORMING" INSTRUCTIONS ON CREATIVE PROBLEM SOLVING BY TRAINED AND UNTRAINED SUBJECTS[1]

SIDNEY J. PARNES AND ARNOLD MEADOW
University of Buffalo

Two initial papers have reported the results of an investigation of the effects of the brainstorming method on creative thinking (Meadow & Parnes, 1959; Meadow, Parnes, & Reese, in press). In this method, a subject is instructed to attempt to solve problems by recording all tentative solutions which occur to him, postponing judicial evaluation of those solutions to a subsequent time period (Osborn, 1957).

The results of the first experiment indicated that a class group instructed for one semester in creative problem solving methods (which emphasized brainstorming) increased its productivity in five of seven tests of creative ability as compared with a control group (Meadow & Parnes, 1959). The second study investigated the effects (in two creative ability problems) of the brainstorming method as compared with a "produce good ideas, penalty for bad ideas" instruction, utilizing the same subjects for each instruction. Results again indicated a superior performance for the brainstorming instructions (Meadow et al., in press). Both of these experiments employed subjects who were taking the course in creative problem solving which emphasized brainstorming methods (Parnes, 1958).

The present study was designed to determine whether the brainstorming instructions would be effective with untrained students who had never received previous instruction in the method. Three hypotheses were formulated: (*a*) With a group of untrained subjects, the brainstorming as compared with nonbrainstorming instructions will produce a significant increment in absolute number of good quality ideas on two creative ability problems. (*b*) The interaction effects found in a previous experiment (Meadow et al., in press) will be confirmed. Specifically, a nonbrainstorming instruction administered first will inhibit the absolute number of good quality ideas produced under a brainstorming instruction which follows. Similarly, a difficult problem administered first will inhibit the absolute number of good quality ideas produced in a relatively easy problem presented in a test period which follows. (*c*) If the brainstorming instructions are administered to a group of students who have taken a one-semester course in creative problem solving emphasizing this technique, and to a group of students who have not taken the course, the former group will produce a significantly greater absolute number of good quality ideas than the latter group.

METHOD

Subjects. In order to test the first two hypotheses, 52 University of Buffalo undergraduate students were used as *S*s. No *S* had taken the creative problem solving course or had received any training in the brainstorming method.

The third hypothesis was tested by comparing a group of University of Buffalo undergraduate creative problem solving course students with a group of students who had not taken the course. Members of the two groups were individually matched on the basis of the variables of grade point average, age, and sex. The matched groups each comprised 17 students selected from a larger group of 21 creative problem solving course students and 26 noncourse students.

Grade point average as based on 12 cred-

[1] This research was financed by a grant from the Creative Education Foundation.

its of courses averaged 1.21, with a standard deviation of .50 for the course group, and 1.21 with a standard deviation of .45 for the noncourse group. Each individual experimental *S* was closely matched on the grade point average variable with each control *S*. The mean of the differences between each of the matched pairs was .13. The average age of the experimental group was 25.50, with a standard deviation of 9.27; average age of the control group was 24.06, with a standard deviation of 6.80. A *t* test indicated no significant difference between these means. The individual experimental versus control *S* matching was less precise on the age variable. The mean of the differences in age for the matched pairs was 6.87 years. There were 11 males and 6 females in the course group, 13 males and 4 females in the control group. The pairs were not closely matched on the sex variable; there were 9 same sex pairs and 8 different sex pairs. This distribution in the matching resulted from the a priori consideration that the grade point average was a more crucial variable than that of age or sex. Close similarity of matched pairs with respect to these latter variables was accordingly sacrificed to achieve greater similarity on the grade point average variable.

The course students differed from the noncourse students in having been exposed to a one-semester course in creative problem solving. The principles studied in this course are those described in Osborn's textbook (1957). An integral part of the teaching method is the repeated practice given to students in the utilization of brainstorming. A specific example of instructions for employing the method is presented in the Procedure section.

Creative problem solving course students were tested during the last two weeks of the semester.

Experimental problems. The Broom and Hanger problems used in the present experiment were selected from Part V of the AC Test of Creative Ability. This test is reported to have differentiated groups of creative from noncreative *S*s (Harris & Simberg, 1954). Data are available from a previous experiment which indicated that scores on the Hanger problem are positively correlated with other tests which have been designed to measure creative thinking (Meadow & Parnes, 1959).[2] Instructions for the Hanger problem required *S*s to list all possible uses for an ordinary wire coat hanger. Similar instructions were given for the Broom problem.

The Examiner's Manual for the AC Test of Creative Ability describes two different scores for Part V, containing the Hanger and Broom problems—a "quantity" score and a "uniqueness" score. If an *S* is instructed to produce ideas without regard for quality (as under the brainstorming instruction), he will generally produce a larger absolute number of ideas than an *S* who either implicitly or explicitly is concerned about the quality of ideas. The crucial question with respect to evaluating the brainstorming method is its efficacy in producing an increment in the number of creative ideas of good quality.

The decision was accordingly made to devise a new scoring method for the test, for the purpose of deriving a reliable "creativity quality" score. This score was based on two attributes, the first of which was uniqueness of the idea. Uniqueness of ideation would seem to represent, however, a necessary but not sufficient condition for creativity. Accordingly, the concept of creativity employed in the present experiment was defined to include also a second attribute—value. This latter attribute was broadly interpreted to include any socially useful value whether it was economic, social, or aesthetic in nature.

[2] Scores on the Hanger problem in the previous experiment correlated with other creative ability tests as follows: Guilford's Unusual Uses, .473; Guilford's Plot Titles High, .452; Guilford's Apparatus, .301; TAT Originality, .520. All but the Apparatus correlation were at the .01 level of significance. The correlation with the Apparatus test was significant at the .05 level.

In implementing this scoring procedure, each response was copied on a separate slip of paper, given a code number, and then presented to the rater for evaluation. This procedure was necessary because *Ss* invariably give a larger *number* of ideas in response to brainstorming than to nonbrainstorming instructions. If the rater scores the responses on the original papers written by *Ss*, he may consequently become aware of the experimental conditions under which the responses were produced. Presentation of each response on a separate coded slip of paper eliminates this type of possible bias.

The rater was instructed to evaluate each response by two separate criteria: (*a*) uniqueness—the degree to which the response deviated from the conventional use of the object, and (*b*) value—the degree to which the response was judged to have social, economic, aesthetic, or other usefulness. The uniqueness attribute was rated on a three-point scale; one point indicated little or no uniqueness, two points indicated moderate uniqueness, three points indicated marked uniqueness. Similarly, the value attribute was rated on the three-point scale of little or no value, moderate value, and marked value. For purposes of testing the hypotheses, these scores were combined into a two-value scale—"good" responses and "bad" responses. A response was scored as "good" if it represented a combined score of at least 5; i.e., it had to be at least moderately valuable and markedly unique, or moderately unique and markedly valuable.

After the responses were individually scored, they were decoded and a total "good response" score assigned to each *S*. Any response which duplicated in essential meaning any other response given by the same *S* was eliminated from the scoring.

The interrater reliabilities for the Hanger and Broom problems, as determined in previous experiments, were .74 and .91, respectively (Meadow & Parnes, 1959; Meadow et al., in press).

Procedure. To test the first two hypotheses, the untrained *Ss* were given first one problem and then the second immediately thereafter. All tests were group administered to each of the four sections separately. One problem was given under brainstorming instructions and the other under nonbrainstorming instructions. The brainstorming instructions were as follows:

> You are to list all the ideas that come to your mind without judging them in any way. Forget about the quality of the ideas entirely. We will count only quantity on this task. Express any idea which comes to your mind. As you go along, you may combine or modify any of the ideas which you have already listed, in order to produce additional ideas. Remember that quantity and freedom of expression without evaluation are the key points.

The corresponding nonbrainstorming instructions were as follows:

> You are to list all the *good* ideas you can think up. Your score will be the total number of good ideas. Don't put down any idea unless you feel it is a *good* one.

The *Ss* were allowed five minutes for each problem.

Half of the *Ss* were given first the Hanger problem and then the Broom problem; the other half were given the problems in the reverse order. Similarly, half of the *Ss* were given first the brainstorming instructions and then the nonbrainstorming instructions, and the other half were given the instructions in the reverse order. Thus, Group 1 was given the Hanger problem under brainstorming instructions as its first problem, and the Broom problem under nonbrainstorming instructions as its second problem; Group 2 was given the Broom problem under brainstorming instructions as its first problem and the Hanger under nonbrainstorming instructions as its second problem. Group 3 was given the Hanger under nonbrainstorming instructions as its first problem and the Broom under brainstorming instructions as its second problem. Finally, Group 4 was given the Broom under nonbrainstorming instructions as its first problem, and the

TABLE 1

MEAN NUMBER OF "GOOD" SOLUTIONS FOR BRAINSTORMING AND NONBRAINSTORMING INSTRUCTIONS BY ADMINISTRATIONS, FOR HANGER AND BROOM PROBLEMS COMBINED

	First Administration			Second Administration			Both Administrations		
	N	$\bar{x}$	σ	*N*	$\bar{x}$	σ	*N*	$\bar{x}$	σ
Instructions									
B	26	3.73	2.86	26	4.81	3.21	52	4.27	3.06
NB	26	3.04	2.03	26	2.00	1.67	52	2.52	1.92

Note:—B refers to brainstorming; NB refers to nonbrainstorming.

TABLE 2

MEAN NUMBER OF "GOOD" SOLUTIONS FOR HANGER AND BROOM PROBLEMS BY ADMINISTRATION, FOR BRAINSTORMING AND NONBRAINSTORMING INSTRUCTIONS COMBINED

	First Administration			Second Administration			Both Administrations		
	N	$\bar{x}$	σ	*N*	$\bar{x}$	σ	*N*	$\bar{x}$	σ
Problems									
Hanger	26	3.77	2.58	26	3.81	3.15	52	3.79	2.85
Broom	26	3.00	2.37	26	3.00	2.64	52	3.00	2.48

TABLE 3

SUMMARY OF ANALYSIS OF VARIANCE OF ABSOLUTE NUMBER OF "GOOD" SOLUTIONS, BRAINSTORMING VS. NONBRAINSTORMING INSTRUCTIONS BY TYPE OF PROBLEM, BY ADMINISTRATIONS

Source	*df*	*MS*	*F*	*P*
Between *Ss*	51			
I × P	1	5.08	<1.00	ns
I × T	1	29.08	3.01	ns
P × T	1	.01	<1.00	ns
error (b)	48	9.67		
Within *Ss*				
Instructions (I)	1	79.63	26.81	<.001
Problems (P)	1	16.17	5.44	<.025
Test Periods (T)	1	.01	<1.00	ns
I × P × T	1	8.09	2.72	ns
error (w)	48	2.97		
Total	103			

Note:—Instructions refers to Brainstorming vs. Nonbrainstorming; Problems refers to Hanger vs. Broom; Test Periods refers to the first test administration vs. the second administration.

Hanger under brainstorming instructions as its second problem.

Lindquist's (1953) Type V analysis of variance was used for statistical analysis. The three experimental variables, Instructions, Problems, and Test Periods (first and second), were all within-*Ss* factors.

In testing the third hypothesis of the experiment, the number of "good" ideas on the Hanger problem under brainstorming instructions for course and noncourse groups was compared.

All tests were administered in groups during regular class hours.

Results

The mean numbers of "good" solutions produced under the two instructions, for the first and second test periods separately, are presented in Table 1. These data show that more "good" ideas were produced under the brainstorming instructions than under the nonbrainstorming instructions, and that this effect was greater in the second test period than in the first. Table 2 indicates that more "good" solutions were produced for the Hanger than for the Broom problem. Evaluation of these differences is shown by the results of the analysis of variance presented in Table 3. The table indicates that the main effects of instructions and problems are significant; the main effect of test periods and all interactions are not significant.

The mean number of "good" ideas on the Hanger problem for course and noncourse students is presented in Table 4. Inspection of the Table indicates a greater average number of "good" ideas for the course group. The difference between these means is significant at the .01 level.

The correlations between total quantity of ideas and "good" ideas are presented in Table 5. The Table indicates positive correlations ranging from .64 to .81 for the two brainstorming and two nonbrainstorming groups of *Ss*. All correlations are significant at the .01 level.

TABLE 4

Comparison of Mean Number of "Good" Solutions of Matched Course and Noncourse Groups under Brainstorming Instructions (Hanger Problem)

	N	$\bar{x}$	σ	t	P
Course Group	17	10.24	3.01	4.95	$<.01$
Noncourse group	17	5.29	3.08		

TABLE 5

Pearson Product-Moment Correlations between Total Quantity of Ideas and Number of "Good" Ideas

	N	r	P
Brainstorming hanger	26	.67	$<.01$
Brainstorming broom	26	.71	$<.01$
Nonbrainstorming hanger	26	.64	$<.01$
Nonbrainstorming broom	26	.81	$<.01$

Discussion and Summary

The experiment was designed to study the effects on creative problem solving (by untrained subjects) of instructions to express solutions without evaluation (brainstorming) and instructions which required only solutions of good quality (nonbrainstorming). The design also allowed for a study of the effects of training in a creative problem solving course emphasizing brainstorming. Each group of *Ss*, (one group untrained, the other trained) was given two problems designed to measure creative ability in two testing periods. One problem was administered under brainstorming instructions; the other problem was administered under nonbrainstorming instructions. The quality of the solutions was later evaluated by a trained rater.

The major findings were as follows: (*a*) Significantly more good quality ideas were produced under brainstorming instructions than under nonbrainstorming instructions. (*b*) The *Ss* trained in a creative problem solving course emphasizing brainstorming

produced a significantly greater number of good quality ideas when using the technique than did the untrained students.

An additional result yielded by the analysis of data indicated a positive correlation between quantity and quality of ideas. This correlation suggests that the efficacy of brainstorming in producing an increment in good ideas is possibly the result of the increased quantity of ideas encouraged by the method—at least in a cognitive problem of the type used in the present experiment. It is likely that in the customary course of daily thinking, some ideas are inhibited by individuals because of fear of criticism from self or others. The brainstorming instruction, in reducing this inhibitory factor and encouraging a greater quantity of ideas, seems to increase also the number of good quality ideas produced. The findings are interpreted to indicate that the brainstorming instruction is an effective method for increasing the production of good ideas in a particular type of creative thinking problem, and that it is even more effective if preceded by extensive training in its use.

REFERENCES

HARRIS, R. R., & SIMBERG, A. L. *AC test of creative ability. (Examiner's manual).* Detroit: General Motors Corp., AC Spark Plug Division, 1954.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

MEADOW, A., & PARNES, S. J. Evaluation of training in creative problem solving. *J. appl. Psychol.*, 1959, **43**, 189–194.

MEADOW, A., PARNES, S. J., & REESE, H. Influence of instructions and problem sequence on a creative problem solving test. *J. appl. Psychol.*, in press.

OSBORN, A. F. *Applied imagination.* New York: Scribner's, 1957.

PARNES, S. J. Description of the University of Buffalo Creative Problem Solving Course. Creative Education Office, Univer. of Buffalo, 1958. (Mimeo.)

Received December 17, 1958.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 50, No. 4, 1959

TRADITIONAL JEWISH CULTURAL VALUES AND PERFORMANCE ON THE WECHSLER TESTS[1]

BORIS M. LEVINSON

Graduate School of Education, Yeshiva University

The purpose of this paper is to suggest that differences between verbal and performance test scores, as well as test scatter, may be attributed to subcultural values emphasizing either verbal or performance skills. These cultural emphases are reflected in the school which in its curriculum extolls and endows one activity with more prestige, thus making another activity appear by comparison less desirable and participation in it ego deflating. As a case in point, the scatter on WISC (Wechsler, 1949) and WAIS (Wechsler, 1955) of Jewish preschool children, elementary school pupils, and college students of traditional background will be discussed.

The Setting

To the traditional Jew, intensive Jewish education is the preferred means of maintaining Judaism. The day (Yeshiva) schools have thus been established to perpetuate not only the religious identity but, also, the national identity (Levinson, 1957). In these schools, very great stress is placed upon verbal ability to the relative neglect of performance arts. A double program is maintained: one in English and one in Hebrew. There is a long school day coupled with additional hours of homework. There is little time left for extracurricular activities, or for physical recreation, or for hobbies not directly related to scholastic pursuits (Levinson, 1959b).

Previous Research

It is well known that environmental pressure will tend to develop certain selective aspects of intellectual ability (Davis, 1948; Eells, Davis, Havighurst, Herrick, Tyler, 1951; Green, 1953; Levinson, 1958a). It is further hypothesized by Hebb (1949) and Piaget (1947) that early perceptual experiences influence intellectual development. The pattern of successes and failures on verbal and performance tests also tends to vary in relation to *cultural needs* (Green, 1953; Strauss, 1954). Dennis (1957), for example, found in the study of the Good-enough Draw-a-Man test that of a group of children from the same socioeconomic background, some were prevented from scoring as high as others due to certain cultural handicaps. The writer (1957) has shown that the intelligence of Jewish children from traditional backgrounds is above the average of the general population. It also appears from certain studies that mental traits become more and more diversified as the child grows older: boys, aged, 9, 12, 15 having the intercorrelation of memory, verbal, and number abilities, decline as follows: .30, .21 and .18 (Garrett, 1946). Further, we find that children from upper socioeconomic levels secure a higher rating on verbal tests than those who come from lower socioeconomic levels (Eells et al., 1951). Attendance at school and, particularly, at college, further enhances the differentiation of intellectual abilities (Hartson, 1936; Levinson, 1958b; Shuey, 1948). The area of specialization at college seems to foster the development of some intellectual traits to the relative neglect of others (Hartson, 1936).

PROCEDURE

Tests Used

The WISC (Wechsler, 1949) and WAIS (Wechsler, 1955) were used in this study. The WISC (Wechsler, 1949) was adminis-

[1] A paper presented in a different form at V Inter-American Congress in Psychology, held in Mexico City, December 1957.
The writer wishes to acknowledge the assistance of Anne Rosenbaum, who has tabulated some of the data.

tered to the preschool and elementary school children and the WAIS (Wechsler, 1955) to college students.

The WISC (Wechsler, 1949) and WAIS (Wechsler, 1955) have both performance and verbal tests and have been standardized on a representative sample of the population. As a matter of fact, Wechsler (1949, p. 1) feels that the materials of the two tests overlap and that the WISC serves the same function in testing children as Wechsler-Bellevue (the precursor of the Wechsler Adult Scale) serves in testing adults (Wechsler, 1944). However, the writer does not wish to imply that the same test items have similar psychological meaning or value for children and adults.

Subjects

The three groups compared were preschool children, elementary school pupils, and college students. Only those *S*s were selected who could be considered typical products of the cultural mold exerted by traditional Jewish values. All the *S*s chosen were "normal" individuals who were not known to have personality difficulties which would require psychotherapy.

Since the WAIS (Levinson, 1958b) study was completed at the time this research was begun, it was necessary to match the preschool and elementary school *S*s to meet the specifications of the WAIS sample. We did not have enough preschool children with the requisite IQ distribution to match the WAIS sample. We finally matched 57 WISC preschool records with 57 WAIS records so selected that the mean and IQ ranges of the 57 WAIS scores corresponded to the mean and IQ of the entire WAIS sample. The preschool sample selected consisted of 41 boys and 16 girls whose average Revised Stanford Binet IQ was 125.20, with an *SD* of 14.00 and whose average age was 5.6 years. These were native born children who were candidates for admission to Jewish traditional day schools and who came from traditional homes. The full scale WISC of these children was 117.05, with an *SD* of 11.10.

The elementary school boys were matched by pairs with the WAIS sample. These were 64 boys whose average age was 11.31 and who were selected from a sample of 122 children who had attended day (Yeshiva) schools since the first grade. The full scale WISC of these pupils was 125.08, with an *SD* of 8.85.

The Yeshiva University *S*s consisted of 64 male volunteer undergraduate and graduate students whose mean age was 21.43. All these students had good command of English and were graduates of elementary and high school Yeshiva schools. The full scale WAIS of these students was 125.08, with an *SD* of 8.85.

Certain limitations regarding the sample must be borne in mind. While the WISC and WAIS full scale IQs were held constant, no allowance was made for possible socioeconomic differences and the possibility (which was not a probability) that the samples chosen were not fully representative of traditional Jewish youth. Further, a question may be raised as to whether children with high performance and low verbal ability who could not compete academically were dropped by the wayside, thus biasing the elementary school and college sample. The writer believes that this selective factor was not operating in this study for two reasons: (*a*) our preschool sample had high verbal ability, its Revised Stanford Binet IQ being 125.20; and (*b*) in the sampled Jewish day schools, with its population preselected on the basis of intelligence, scholastic achievement depends largely on attitude, motivation, etc.

Results

Table 1 presents the means and *t*'s for the preschool, elementary school, and University *S*s. We may note that while the difference between the verbal and performance IQs for the preschool children is significant only at the .10 level of confidence, the differences for the elementary and University *S*s are at the .001 level of confidence.

We may further note that when the

TABLE 1

SIGNIFICANCE OF THE DIFFERENCES BETWEEN THE MEANS OF THE VERBAL AND PERFORMANCE IQS OF THE PRESCHOOL, ELEMENTARY SCHOOL, AND UNIVERSITY STUDENTS

	Verbal Scale	Perf. Scale	Diff.	*t*
Preschool	117.19	114.05	3.14	1.83*
Elementary School	125.08	105.27	19.81	10.97**
University	125.59	105.30	20.29	10.75**

* .10 level of significance.
** .01 level of significance.

differences between the differences shown in Table 1 are compared, those involving the preschool and elementary school and for the preschool and University comparisons are significant at the .01 level of confidence (Walker & Lev, 1953, p. 158). However, the difference between the differences in the instance of the elementary school and University *Ss* is statistically insignificant. These *t*'s are not shown in the table.

Table 2 shows the means and *SD*s of the scaled score equivalents of the subtests of the verbal and of the performance parts of (*a*) WISC for the preschool and elementary school children and (*b*) WAIS for the University students. An analysis of the table indicates: (*a*) minimal scatter among the subtests of the preschool children, the mean subtest scores ranging from 11.36 to 13.04, (*b*) pronounced scatter among elementary school pupils and college students, elevated verbal subtests and depressed performance subtest scores being characteristic of both. The mean scores of the elementary school children range from 10.16 to 15.16 and of the University students from 8.84 to 15.64.

DISCUSSION

The writer is of the opinion that the cultural pressure exerted towards verbal learning has brought about this differentiation of ability at the elementary and college levels. He further thinks that somewhat similar psychometric patterns may be found whenever there is great emphasis on verbal learning and a relative neglect of performance arts. The writer (in press [b]) found such patterns among Irish, Italian

TABLE 2

MEANS AND *SD*S OF THE SCALED SCORE EQUIVALENTS OF THE WISC AND THE WAIS

	Preschool		Elementary School		University	
	Mean	*SD*	Mean	*SD*	Mean	*SD*
Verbal[a]						
Information	12.80	3.42	14.00	2.56	15.64	1.95
Comprehension	12.98	2.70	15.16	2.90	15.49	2.43
Arithmetic	13.04	2.90	13.59	2.54	13.84	2.50
Similarities	12.40	3.24	14.00	2.52	13.20	2.58
Vocabulary	11.95	3.52	13.94	2.46	14.48	2.38
Performance						
Picture Completion	12.38	2.20	10.84	3.00	11.75	2.40
Picture Arrangement	12.34	2.14	11.00	2.82	10.30	2.09
Block Design	12.62	2.30	10.37	2.74	11.02	2.60
Object Assembly	11.36	2.46	10.16	2.80	8.84	2.65
Coding (WISC) or Digit Symbol (WAIS)	11.50	3.12	11.42	3.10	12.02	2.88
N[b]	57		64		64	

[a] Digit Span omitted.
[b] *N*s are as given, except in *Preschool* (Similarities and Object Assembly $N = 56$; Vocabulary $N = 55$; Coding $N = 53$) and in *Elementary School* (Coding $N = 63$).

and Jewish children who were attending parochial schools. The verbal scores were higher than the performance scores in accordance with the cultural pressures exerted for verbal accomplishment by the various subcultures and their schools. As the writer noted in another connection (Levinson, 1958b), there is no a priori reason why a person with a superior intelligence should have a higher verbal than performance ability, providing, of course, that both verbal and performance items are equally valid tests of intelligence. The importance of the subculture from which the individual comes with its emphasis on verbal skills and the premium on verbal accomplishment is often overlooked.

The general implication that we draw is that the greater the premium placed by a certain subculture on language and abstract thinking, the greater will be the disparity between verbal and performance abilities.

This raises certain questions with respect to the Wechsler tests, and presumably other tests. Is it possible, for example, that the subtests, which are considered good measures (Wechsler, 1944) of important functions in the general population, do not have the same validity when applied to *Ss* from a highly verbal Jewish traditional culture?

Summary

A comparative study was made of the performance and verbal abilities of preschool children, elementary school pupils, and college students of traditional Jewish background. The groups were matched for full scale IQs on either WISC or WAIS. The difference between the verbal and performance IQs for the preschool children was at the .10 level of confidence. However, on the elementary and college levels, differences were found which were statistically significant at the .001 level. This is interpreted as being due to the cumulative effect of Jewish cultural values emphasizing verbal abilities.

REFERENCES

Davis, W. A. *Social class influences upon learning.* Cambridge, Mass.: Harvard Univer. Press, 1948.

Dennis, W. Performance of Near Eastern children on the Draw-a-Man test. *Child Developm.*, 1957, **28**, 427–431.

Eells, K., Davis, A., Havighurst, R. J., Herrick, V. E., & Tyler, R. W. *Intelligence and cultural differences.* Chicago: Univer. Chicago Press, 1951.

Garrett, H. E. A developmental theory of intelligence. *Amer. Psychologist*, 1946, **1**, 372–378.

Green, T. C. Individual response to cultural determinants. *J. educ. Sociol.*, 1953, **26**, 392–399.

Hartson, L. Does college training influence test intelligence? *J. educ. Psychol.*, 1936, **27**, 481–491.

Hebb, D. O. *The organization of behavior.* New York: Wiley, 1949.

Levinson, B. M. The intelligence of applicants for admission to Jewish day schools. *Jewish soc. Stud.*, 1957, **19**, 129–140.

Levinson, B. M. Culture and mental retardation. *Psychol. Rec.*, 1958, **8**, 27–38. (a)

Levinson, B. M. Cultural pressure and WAIS scatter in a traditional Jewish setting. *J. genet. Psychol.*, 1958, **93**, 277–286. (b)

Levinson, B. M. The problems of Jewish religious youth. *Genet. psychol. Monogr.*, in press. (a)

Levinson, B. M. Subcultural variations in performance and verbal ability at the elementary school level. *J. genet. Psychol.*, in press. (b)

Piaget, J. *The psychology of intelligence.* New York: Harcourt, 1947.

Shuey, A. M. Improvement in scores on American Council Psychological Examination from freshman to senior year. *J. educ. Psychol.*, 1948, **39**, 417–426.

Strauss, M. Subcultural variation in Ceylonese mental ability: A study of national character. *J. soc. Psychol.*, 1954, **39**, 129–141.

WALKER, HELEN, M., & LEV, J. *Statistical inference*. New York: Holt, 1953.

WECHSLER, D. *The measurement of adult intelligence*. (3rd ed.) Baltimore: Williams & Wilkins, 1944.

WECHSLER, D. *Wechsler Intelligence Scale for Children*. New York: Psychological Corp., 1949.

WECHSLER, D. *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corp., 1955.

Received December 20, 1958.

MICROFORM

All journals published by the American Psychological Association are being made available on MICROFILM or MICROCARD.

Psychological Review
American Psychologist
Psychological Bulletin
Psychological Abstracts
Contemporary Psychology
Psychological Monographs
Journal of Applied Psychology
Journal of Consulting Psychology
Journal of Educational Psychology
Journal of Experimental Psychology
Journal of Abnormal and Social Psychology
Journal of Comparative and Physiological Psychology

Available only in volume units; no single issues.

For MICROFILM, order from:

UNIVERSITY MICROFILMS, INC.
313 North First Street
Ann Arbor
Michigan

For MICROCARD, order from:

J. S. CANNER & COMPANY, INC.
Microcard Division
618 Parker Street
Roxbury 20, Massachusetts

Again Available

SCHOOL PSYCHOLOGISTS AT MID-CENTURY

Edited by

NORMA E. CUTTS

The Report of the Thayer Conference, published in 1955, covered an intensive study of the development of school psychology. Because of continuing demand, the Report has been reprinted.

230 pages. Price $2.75

AMERICAN PSYCHOLOGICAL ASSOCIATION
1333 Sixteenth Street, N.W.
Washington 6, D. C.

The
Journal
of
Educational
Psychology